

# **METHODS FOR DETECTING DELETIONS AND DUPLICATIONS OF CHROMOSOME SEGMENTS**

## **FIELD OF THE INVENTION**

[1] The present invention generally relates to methods of detecting deletions and duplications of chromosome segments or entire chromosomes.

## **BACKGROUND OF THE INVENTION**

[2] Copy number variation (CNV) has been identified as a major cause of structural variation in the genome, involving both duplications and deletions of sequences that typically range in length from 1,000 base pairs (1 kb) to 20 megabases (mb). Deletions and duplications of chromosome segments or entire chromosomes are associated with a variety of conditions, such as susceptibility or resistance to disease.

[3] CNVs are often assigned to one of two main categories, based on the length of the affected sequence. The first category includes copy number polymorphisms (CNPs), which are common in the general population, occurring with an overall frequency of greater than 1%. CNPs are typically small (most are less than 10 kilobases in length), and they are often enriched for genes that encode proteins important in drug detoxification and immunity. A subset of these CNPs is highly variable with respect to copy number. As a result, different human chromosomes can have a wide range of copy numbers (*e.g.*, 2, 3, 4, 5, *etc.*) for a particular set of genes. CNPs associated with immune response genes have recently been associated with susceptibility to complex genetic diseases, including psoriasis, Crohn's disease, and glomerulonephritis.

[4] The second class of CNVs includes relatively rare variants that are much longer than CNPs, ranging in size from hundreds of thousands of base pairs to over 1 million base pairs in length. In some cases, these CNVs may have arisen during production of the sperm or egg that gave rise to a particular individual, or they may have been passed down for only a few generations within a family. These large and rare structural variants have been observed disproportionately in subjects with mental retardation, developmental delay, schizophrenia, and autism. Their appearance in such subjects has led to speculation that large and rare CNVs may be more important in neurocognitive diseases than other forms of inherited mutations, including single nucleotide substitutions.

[5] Gene copy number can be altered in cancer cells. For instance, duplication of *Chrlp* is common in breast cancer, and the *EGFR* copy number can be higher than normal in non-small cell lung cancer. Cancer is one of the leading causes of death; thus, early diagnosis and treatment of cancer is important, since it

# METHODS FOR DETECTING DELETIONS AND DUPLICATIONS OF CHROMOSOME SEGMENTS

## FIELD OF THE INVENTION

[1] The present invention generally relates to methods of detecting deletions and duplications of chromosome segments or entire chromosomes.

## BACKGROUND OF THE INVENTION

[2] Copy number variation (CNV) has been identified as a major cause of structural variation in the genome, involving both duplications and deletions of sequences that typically range in length from 1,000 base pairs (1 kb) to 20 megabases (mb). Deletions and duplications of chromosome segments or entire chromosomes are associated with a variety of conditions, such as susceptibility or resistance to disease.

[3] CNVs are often assigned to one of two main categories, based on the length of the affected sequence. The first category includes copy number polymorphisms (CNPs), which are common in the general population, occurring with an overall frequency of greater than 1%. CNPs are typically small (most are less than 10 kilobases in length), and they are often enriched for genes that encode proteins important in drug detoxification and immunity. A subset of these CNPs is highly variable with respect to copy number. As a result, different human chromosomes can have a wide range of copy numbers (*e.g.*, 2, 3, 4, 5, *etc.*) for a particular set of genes. CNPs associated with immune response genes have recently been associated with susceptibility to complex genetic diseases, including psoriasis, Crohn's disease, and glomerulonephritis.

[4] The second class of CNVs includes relatively rare variants that are much longer than CNPs, ranging in size from hundreds of thousands of base pairs to over 1 million base pairs in length. In some cases, these CNVs may have arisen during production of the sperm or egg that gave rise to a particular individual, or they may have been passed down for only a few generations within a family. These large and rare structural variants have been observed disproportionately in subjects with mental retardation, developmental delay, schizophrenia, and autism. Their appearance in such subjects has led to speculation that large and rare CNVs may be more important in neurocognitive diseases than other forms of inherited mutations, including single nucleotide substitutions.

[5] Gene copy number can be altered in cancer cells. For instance, duplication of *Chrlp* is common in breast cancer, and the *EGFR* copy number can be higher than normal in non-small cell lung cancer. Cancer is one of the leading causes of death; thus, early diagnosis and treatment of cancer is important, since it

can improve the patient's outcome (such as by increasing the probability of remission and the duration of remission). Early diagnosis can also allow the patient to undergo fewer or less drastic treatment alternatives. Many of the current treatments that destroy cancerous cells also affect normal cells, resulting in a variety of possible side-effects, such as nausea, vomiting, low blood cell counts, increased risk of infection, hair loss, and ulcers in mucous membranes. Thus, early detection of cancer is desirable since it can reduce the amount and/or number of treatments (such as chemotherapeutic agents or radiation) needed to eliminate the cancer.

[6] Copy number variation has also been associated with severe mental and physical handicaps, and idiopathic learning disability. Non-invasive prenatal testing (NIPT) using cell-free DNA (cfDNA) can be used to detect abnormalities, such as fetal trisomies 13, 18, and 21, triploidy, and sex chromosome aneuploidies. Subchromosomal microdeletions, which can also result in severe mental and physical handicaps, are more challenging to detect due to their smaller size. Eight of the microdeletion syndromes have an aggregate incidence of more than 1 in 1000, making them nearly as common as fetal autosomal trisomies.

[7] In addition, a higher copy number of CCL3L1 has been associated with lower susceptibility to HIV infection, and a low copy number of FCGR3B (the CD 16 cell surface immunoglobulin receptor) can increase susceptibility to systemic lupus erythematosus and similar inflammatory autoimmune disorders.

[8] Thus, improved methods are needed to detect deletions and duplications of chromosome segments or entire chromosomes. Preferably, these methods can be used to more accurately diagnose disease or an increased risk of disease, such as cancer or CNVs in a gestating fetus.

## **SUMMARY OF THE INVENTION**

[9] In one aspect, the invention features a method of determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of one or more cells from an individual. In some embodiments, the method includes obtaining phased genetic data for the first homologous chromosome segment comprising, the identity of the allele present at that locus on the first homologous chromosome segment for each locus in a set of polymorphic loci on the first homologous chromosome segment, obtaining phased genetic data for the second homologous chromosome segment comprising the identity of the allele present at that locus on the second homologous chromosome segment for each locus in the set of polymorphic loci on the second homologous chromosome segment, and obtaining measured genetic allelic data comprising the

can improve the patient's outcome (such as by increasing the probability of remission and the duration of remission). Early diagnosis can also allow the patient to undergo fewer or less drastic treatment alternatives. Many of the current treatments that destroy cancerous cells also affect normal cells, resulting in a variety of possible side-effects, such as nausea, vomiting, low blood cell counts, increased risk of infection, hair loss, and ulcers in mucous membranes. Thus, early detection of cancer is desirable since it can reduce the amount and/or number of treatments (such as chemotherapeutic agents or radiation) needed to eliminate the cancer.

**[6]** Copy number variation has also been associated with severe mental and physical handicaps, and idiopathic learning disability. Non-invasive prenatal testing (NIPT) using cell-free DNA (cfDNA) can be used to detect abnormalities, such as fetal trisomies 13, 18, and 21, triploidy, and sex chromosome aneuploidies. Subchromosomal microdeletions, which can also result in severe mental and physical handicaps, are more challenging to detect due to their smaller size. Eight of the microdeletion syndromes have an aggregate incidence of more than 1 in 1000, making them nearly as common as fetal autosomal trisomies.

**[7]** In addition, a higher copy number of CCL3L1 has been associated with lower susceptibility to HIV infection, and a low copy number of FCGR3B (the CD 16 cell surface immunoglobulin receptor) can increase susceptibility to systemic lupus erythematosus and similar inflammatory autoimmune disorders.

**[8]** Thus, improved methods are needed to detect deletions and duplications of chromosome segments or entire chromosomes. Preferably, these methods can be used to more accurately diagnose disease or an increased risk of disease, such as cancer or CNVs in a gestating fetus.

## **SUMMARY OF THE INVENTION**

**[9]** In one aspect, the invention features a method of determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of one or more cells from an individual. In some embodiments, the method includes obtaining phased genetic data for the first homologous chromosome segment comprising, the identity of the allele present at that locus on the first homologous chromosome segment for each locus in a set of polymorphic loci on the first homologous chromosome segment, obtaining phased genetic data for the second homologous chromosome segment comprising the identity of the allele present at that locus on the second homologous chromosome segment for each locus in the set of polymorphic loci on the second homologous chromosome segment, and obtaining measured genetic allelic data comprising the

amount of each allele present in a sample of DNA or RNA from one or more cells from the individual, for each of the alleles at each of the loci in the set of polymorphic loci. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment in the genome of one or more cells from the individual, calculating (such as calculating on a computer) a likelihood of one or more of the hypotheses based on the obtained genetic data of the sample and the obtained phased genetic data, and selecting the hypothesis with the greatest likelihood, thereby determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more cells from the individual. In some embodiments, the phased data includes inferred phased data using population based haplotype frequencies and/or measured phased data (*e.g.*, phased data obtained by measuring a sample containing DNA or RNA from the individual or a relative of the individual).

**[10]** In one aspect, the invention provides a method for determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of one or more cells from an individual. In some embodiments, the method includes obtaining phased genetic data for the first homologous chromosome segment comprising the identity of the allele present at that locus on the first homologous chromosome segment for each locus in a set of polymorphic loci on the first homologous chromosome segment, obtaining phased genetic data for the second homologous chromosome segment comprising the identity of the allele present at that locus on the second homologous chromosome segment for each locus in the set of polymorphic loci on the second homologous chromosome segment, and obtaining measured genetic allelic data comprising the amount of each allele present in a sample of DNA or RNA from one or more cells from the individual for each of the alleles at each of the loci in the set of polymorphic loci. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment; calculating, for each of the hypotheses, expected genetic data for the plurality of loci in the sample from the obtained phased genetic data; calculating (such as calculating on a computer) the data fit between the obtained genetic data of the sample and the expected genetic data for the sample; ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more cells from the individual.

**[11]** In one aspect, the invention features a method for determining if there is an overrepresentation of

amount of each allele present in a sample of DNA or RNA from one or more cells from the individual, for each of the alleles at each of the loci in the set of polymorphic loci. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment in the genome of one or more cells from the individual, calculating (such as calculating on a computer) a likelihood of one or more of the hypotheses based on the obtained genetic data of the sample and the obtained phased genetic data, and selecting the hypothesis with the greatest likelihood, thereby determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more cells from the individual. In some embodiments, the phased data includes inferred phased data using population based haplotype frequencies and/or measured phased data (*e.g.*, phased data obtained by measuring a sample containing DNA or RNA from the individual or a relative of the individual).

**[10]** In one aspect, the invention provides a method for determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of one or more cells from an individual. In some embodiments, the method includes obtaining phased genetic data for the first homologous chromosome segment comprising the identity of the allele present at that locus on the first homologous chromosome segment for each locus in a set of polymorphic loci on the first homologous chromosome segment, obtaining phased genetic data for the second homologous chromosome segment comprising the identity of the allele present at that locus on the second homologous chromosome segment for each locus in the set of polymorphic loci on the second homologous chromosome segment, and obtaining measured genetic allelic data comprising the amount of each allele present in a sample of DNA or RNA from one or more cells from the individual for each of the alleles at each of the loci in the set of polymorphic loci. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment; calculating, for each of the hypotheses, expected genetic data for the plurality of loci in the sample from the obtained phased genetic data; calculating (such as calculating on a computer) the data fit between the obtained genetic data of the sample and the expected genetic data for the sample; ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more cells from the individual.

**[11]** In one aspect, the invention features a method for determining if there is an overrepresentation of

the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of one or more cells from an individual. In some embodiments, the method includes obtaining phased genetic data for the first homologous chromosome segment comprising the identity of the allele present at that locus on the first homologous chromosome segment for each locus in a set of polymorphic loci on the first homologous chromosome segment, obtaining phased genetic data for the second homologous chromosome segment comprising the identity of the allele present at that locus on the second homologous chromosome segment for each locus in the set of polymorphic loci on the second homologous chromosome segment, and obtaining measured genetic allelic data comprising, for each of the alleles at each of the loci in the set of polymorphic loci, the amount of each allele present in a sample of DNA or RNA from one or more target cells and one or more non-target cells from the individual. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment; calculating (such as calculating on a computer), for each of the hypotheses, expected genetic data for the plurality of loci in the sample from the obtained phased genetic data for one or more possible ratios of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample; calculating (such as calculating on a computer) for each possible ratio of DNA or RNA and for each hypothesis, the data fit between the obtained genetic data of the sample and the expected genetic data for the sample for that possible ratio of DNA or RNA and for that hypothesis; ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more cells from the individual.

**[12]** In one aspect, the invention features a method for determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of one or more cells from an individual. In some embodiments, the method includes obtaining phased genetic data for the first homologous chromosome segment comprising the identity of the allele present at that locus on the first homologous chromosome segment for each locus in a set of polymorphic loci on the first homologous chromosome segment, obtaining phased genetic data for the second homologous chromosome segment comprising the identity of the allele present at that locus on the second homologous chromosome segment for each locus in the set of polymorphic loci on the second homologous chromosome segment, and obtaining measured genetic allelic data comprising the amount of each allele present in a sample of DNA or RNA from one or more target cells and one or more

the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of one or more cells from an individual. In some embodiments, the method includes obtaining phased genetic data for the first homologous chromosome segment comprising the identity of the allele present at that locus on the first homologous chromosome segment for each locus in a set of polymorphic loci on the first homologous chromosome segment, obtaining phased genetic data for the second homologous chromosome segment comprising the identity of the allele present at that locus on the second homologous chromosome segment for each locus in the set of polymorphic loci on the second homologous chromosome segment, and obtaining measured genetic allelic data comprising, for each of the alleles at each of the loci in the set of polymorphic loci, the amount of each allele present in a sample of DNA or RNA from one or more target cells and one or more non-target cells from the individual. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment; calculating (such as calculating on a computer), for each of the hypotheses, expected genetic data for the plurality of loci in the sample from the obtained phased genetic data for one or more possible ratios of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample; calculating (such as calculating on a computer) for each possible ratio of DNA or RNA and for each hypothesis, the data fit between the obtained genetic data of the sample and the expected genetic data for the sample for that possible ratio of DNA or RNA and for that hypothesis; ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more cells from the individual.

**[12]** In one aspect, the invention features a method for determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of one or more cells from an individual. In some embodiments, the method includes obtaining phased genetic data for the first homologous chromosome segment comprising the identity of the allele present at that locus on the first homologous chromosome segment for each locus in a set of polymorphic loci on the first homologous chromosome segment, obtaining phased genetic data for the second homologous chromosome segment comprising the identity of the allele present at that locus on the second homologous chromosome segment for each locus in the set of polymorphic loci on the second homologous chromosome segment, and obtaining measured genetic allelic data comprising the amount of each allele present in a sample of DNA or RNA from one or more target cells and one or more

non-target cells from the individual for each of the alleles at each of the loci in the set of polymorphic loci. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment; calculating (such as calculating on a computer), for each of the hypotheses, expected genetic data for the plurality of loci in the sample from the obtained phased genetic data for one or more possible ratios of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample; calculating (such as calculating on a computer) for each locus in the plurality of loci, each possible ratio of DNA or RNA, and each hypothesis, the likelihood that the hypothesis is correct by comparing the obtained genetic data of the sample for that locus and the expected genetic data for that locus for that possible ratio of DNA or RNA and for that hypothesis; determining the combined probability for each hypothesis by combining the probabilities of that hypothesis for each locus and each possible ratio; and selecting the hypothesis with the greatest combined probability, thereby determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment. In some embodiments, all of the loci are considered at once to calculate the probability of a particular hypothesis, and the hypothesis with the greatest probability is selected.

**[13]** In one aspect, the invention features a method for determining a number of copies of a chromosome segment of interest in the genome of a fetus. In some embodiments, the method includes obtaining phased genetic data for at least one biological parent of the fetus, wherein the phased genetic data comprises the identity of the allele present for each locus in a set of polymorphic loci on a first homologous chromosome segment and a second homologous chromosome segment in a pair of homologous chromosome segments that comprises the chromosome segment of interest. In some embodiments, the method includes obtaining genetic data at the set of polymorphic loci on the chromosome segment of interest in a mixed sample of DNA or RNA comprising fetal DNA or RNA and maternal DNA or RNA from the mother of the fetus by measuring the quantity of each allele at each locus. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the number of copies of the chromosome segment of interest present in the genome of the fetus. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying, for one or both parents, the number of copies of the first homologous chromosome segment or portion thereof from the parent in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion thereof from the parent in the genome of the fetus, and the total number of copies of the chromosome segment of interest present in the genome of the fetus. In some embodiments, the method

non-target cells from the individual for each of the alleles at each of the loci in the set of polymorphic loci. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment; calculating (such as calculating on a computer), for each of the hypotheses, expected genetic data for the plurality of loci in the sample from the obtained phased genetic data for one or more possible ratios of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample; calculating (such as calculating on a computer) for each locus in the plurality of loci, each possible ratio of DNA or RNA, and each hypothesis, the likelihood that the hypothesis is correct by comparing the obtained genetic data of the sample for that locus and the expected genetic data for that locus for that possible ratio of DNA or RNA and for that hypothesis; determining the combined probability for each hypothesis by combining the probabilities of that hypothesis for each locus and each possible ratio; and selecting the hypothesis with the greatest combined probability, thereby determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment. In some embodiments, all of the loci are considered at once to calculate the probability of a particular hypothesis, and the hypothesis with the greatest probability is selected.

**[13]** In one aspect, the invention features a method for determining a number of copies of a chromosome segment of interest in the genome of a fetus. In some embodiments, the method includes obtaining phased genetic data for at least one biological parent of the fetus, wherein the phased genetic data comprises the identity of the allele present for each locus in a set of polymorphic loci on a first homologous chromosome segment and a second homologous chromosome segment in a pair of homologous chromosome segments that comprises the chromosome segment of interest. In some embodiments, the method includes obtaining genetic data at the set of polymorphic loci on the chromosome segment of interest in a mixed sample of DNA or RNA comprising fetal DNA or RNA and maternal DNA or RNA from the mother of the fetus by measuring the quantity of each allele at each locus. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the number of copies of the chromosome segment of interest present in the genome of the fetus. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying, for one or both parents, the number of copies of the first homologous chromosome segment or portion thereof from the parent in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion thereof from the parent in the genome of the fetus, and the total number of copies of the chromosome segment of interest present in the genome of the fetus. In some embodiments, the method

includes calculating (such as calculating on a computer), for each of the hypotheses, expected genetic data for the plurality of loci in the mixed sample from the obtained phased genetic data from the parent(s); calculating (such as calculating on a computer) the data fit between the obtained genetic data of the mixed sample and the expected genetic data for the mixed sample; ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the number of copies of the chromosome segment of interest in the genome of the fetus.

**[14]** In one aspect, the invention features a method for determining a number of copies of a chromosome or chromosome segment of interest in the genome of a fetus. In some embodiments, the method includes obtaining phased genetic data for at least one biological parent of the fetus, wherein the phased genetic data comprises the identity of the allele present for each locus in a set of polymorphic loci on a first homologous chromosome segment and a second homologous chromosome segment in the parent. In some embodiments, the method includes obtaining genetic data at the set of polymorphic loci on the chromosome or chromosome segment in a mixed sample of DNA or RNA comprising fetal DNA or

RNA and maternal DNA or RNA from the mother of the fetus by measuring the quantity of each allele at each locus. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the number of copies of the chromosome or chromosome segment of interest present in the genome of the fetus. In some embodiments, the method includes creating (such as creating on a computer) for each of the hypotheses, a probability distribution of the expected quantity of each allele at each of the plurality of loci in mixed sample from the (i) the obtained phased genetic data from the parent(s) and (ii) optionally the probability of one or more crossovers that may have occurred during the formation of a gamete that contributed a copy of the chromosome or chromosome segment of interest to the fetus; calculating (such as calculating on a computer) a fit, for each of the hypotheses, between (1) the obtained genetic data of the mixed sample and (2) the probability distribution of the expected quantity of each allele at each of the plurality of loci in mixed sample for that hypothesis; ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the number of copies of the chromosome segment of interest in the genome of the fetus.

**[15]** In some embodiments, the method includes obtaining phased genetic data for the mother of the fetus. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the number of copies of the first homologous chromosome segment or portion thereof from the mother in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion

includes calculating (such as calculating on a computer), for each of the hypotheses, expected genetic data for the plurality of loci in the mixed sample from the obtained phased genetic data from the parent(s); calculating (such as calculating on a computer) the data fit between the obtained genetic data of the mixed sample and the expected genetic data for the mixed sample; ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the number of copies of the chromosome segment of interest in the genome of the fetus.

**[14]** In one aspect, the invention features a method for determining a number of copies of a chromosome or chromosome segment of interest in the genome of a fetus. In some embodiments, the method includes obtaining phased genetic data for at least one biological parent of the fetus, wherein the phased genetic data comprises the identity of the allele present for each locus in a set of polymorphic loci on a first homologous chromosome segment and a second homologous chromosome segment in the parent. In some embodiments, the method includes obtaining genetic data at the set of polymorphic loci on the chromosome or chromosome segment in a mixed sample of DNA or RNA comprising fetal DNA or

RNA and maternal DNA or RNA from the mother of the fetus by measuring the quantity of each allele at each locus. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the number of copies of the chromosome or chromosome segment of interest present in the genome of the fetus. In some embodiments, the method includes creating (such as creating on a computer) for each of the hypotheses, a probability distribution of the expected quantity of each allele at each of the plurality of loci in mixed sample from the (i) the obtained phased genetic data from the parent(s) and (ii) optionally the probability of one or more crossovers that may have occurred during the formation of a gamete that contributed a copy of the chromosome or chromosome segment of interest to the fetus; calculating (such as calculating on a computer) a fit, for each of the hypotheses, between (1) the obtained genetic data of the mixed sample and (2) the probability distribution of the expected quantity of each allele at each of the plurality of loci in mixed sample for that hypothesis; ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the number of copies of the chromosome segment of interest in the genome of the fetus.

**[15]** In some embodiments, the method includes obtaining phased genetic data for the mother of the fetus. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the number of copies of the first homologous chromosome segment or portion thereof from the mother in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion

thereof from the mother in the genome of the fetus, and the total number of copies of the chromosome segment of interest present in the genome of the fetus. In some embodiments, the method includes calculating, for each of the hypotheses, expected genetic data for the plurality of loci in the mixed sample from the obtained phased genetic data from the mother.

**[16]** In some embodiments, the expected genetic data for each of the hypotheses comprises the identity and an amount of one or more alleles at each locus in the plurality of loci from the maternal DNA or RNA and fetal DNA or RNA in the mixed sample. In some embodiments, the method includes calculating (such as calculating on a computer) expected genetic data by determining a fraction of fetal DNA or RNA and a fraction of maternal DNA or RNA in the mixed sample. In some embodiments, the method includes calculating, for each locus in the plurality of loci, the expected amount of one or more of the alleles for that locus in the maternal

DNA or RNA in the mixed sample using the identity of the allele(s) present at that locus in the obtained phased genetic data of the mother and the fraction of maternal DNA or RNA in the mixed sample. In some embodiments, the method includes calculating (such as calculating on a computer), for each locus in the plurality of loci for each hypothesis, the expected amount of one or more of the alleles for that locus in the fetal DNA or RNA inherited from the mother in the mixed sample using the identity of the allele present at that locus in the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, the number of copies of the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, and the fraction of fetal DNA or RNA in the mixed sample.

**[17]** In some embodiments, the expected genetic data for each of the hypotheses comprises the identity and an amount of one or more alleles at each locus in the plurality of loci from the maternal DNA or RNA and fetal DNA or RNA in the mixed sample. In some embodiments, the method includes calculating expected genetic data by determining a fraction of fetal DNA or RNA and a fraction of maternal DNA or RNA in the mixed sample. In some embodiments, the method includes calculating (such as calculating on a computer), for each locus in the plurality of loci, the expected amount of one or more of the alleles for that locus in the maternal DNA or RNA in the mixed sample using the identity of the allele(s) present at that locus in the obtained phased genetic data of the mother and the fraction of maternal DNA or RNA in the mixed sample. In some embodiments, the method includes calculating (such as calculating on a computer), for each locus in the plurality of loci for each hypothesis, the expected amount of one or more of the alleles for that locus in the fetal DNA or RNA inherited from the mother in the mixed sample using

thereof from the mother in the genome of the fetus, and the total number of copies of the chromosome segment of interest present in the genome of the fetus. In some embodiments, the method includes calculating, for each of the hypotheses, expected genetic data for the plurality of loci in the mixed sample from the obtained phased genetic data from the mother.

**[16]** In some embodiments, the expected genetic data for each of the hypotheses comprises the identity and an amount of one or more alleles at each locus in the plurality of loci from the maternal DNA or RNA and fetal DNA or RNA in the mixed sample. In some embodiments, the method includes calculating (such as calculating on a computer) expected genetic data by determining a fraction of fetal DNA or RNA and a fraction of maternal DNA or RNA in the mixed sample. In some embodiments, the method includes calculating, for each locus in the plurality of loci, the expected amount of one or more of the alleles for that locus in the maternal

DNA or RNA in the mixed sample using the identity of the allele(s) present at that locus in the obtained phased genetic data of the mother and the fraction of maternal DNA or RNA in the mixed sample. In some embodiments, the method includes calculating (such as calculating on a computer), for each locus in the plurality of loci for each hypothesis, the expected amount of one or more of the alleles for that locus in the fetal DNA or RNA inherited from the mother in the mixed sample using the identity of the allele present at that locus in the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, the number of copies of the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, and the fraction of fetal DNA or RNA in the mixed sample.

**[17]** In some embodiments, the expected genetic data for each of the hypotheses comprises the identity and an amount of one or more alleles at each locus in the plurality of loci from the maternal DNA or RNA and fetal DNA or RNA in the mixed sample. In some embodiments, the method includes calculating expected genetic data by determining a fraction of fetal DNA or RNA and a fraction of maternal DNA or RNA in the mixed sample. In some embodiments, the method includes calculating (such as calculating on a computer), for each locus in the plurality of loci, the expected amount of one or more of the alleles for that locus in the maternal DNA or RNA in the mixed sample using the identity of the allele(s) present at that locus in the obtained phased genetic data of the mother and the fraction of maternal DNA or RNA in the mixed sample. In some embodiments, the method includes calculating (such as calculating on a computer), for each locus in the plurality of loci for each hypothesis, the expected amount of one or more of the alleles for that locus in the fetal DNA or RNA inherited from the mother in the mixed sample using

the identity of the allele present at that locus in the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, the number of copies of the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, the identity of one or more possible alleles at that locus in the first or second homologous chromosome segment from the father that is specified by the hypothesis to have been inherited by the fetus, the number of copies of the first or second homologous chromosome segment from the father that is specified by the hypothesis to have been inherited by the fetus, and the fraction of fetal DNA or RNA in the mixed sample. In some embodiments, population frequencies are used to predict the identity of the alleles in the first or second homologous chromosome segment from the father. In some embodiments, the probability for each of the possible alleles at each locus in the first or second homologous chromosome segment from the father are considered to be the same.

**[18]** In some embodiments, the method includes obtaining phased genetic data for both the mother and father of the fetus. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the number of copies of the first homologous chromosome segment or portion thereof from the mother in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion thereof from the mother in the genome of the fetus, the number of copies of the first homologous chromosome segment or portion thereof from the father in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion thereof from the father in the genome of the fetus, and the total number of copies of the chromosome segment of interest present in the genome of the fetus. In some embodiments, the method includes calculating (such as calculating on a computer), for each of the hypotheses, expected genetic data for the plurality of loci in the mixed sample from the obtained phased genetic data from the mother and obtained phased genetic data from the father.

**[19]** In some embodiments, the expected genetic data for each of the hypotheses comprises the identity and an amount of one or more alleles at each locus in the plurality of loci from the maternal DNA or RNA and fetal DNA or RNA in the mixed sample. In some embodiments, the method includes calculating expected genetic data by determining a fraction of fetal DNA or RNA and a fraction of maternal DNA or RNA in the mixed sample. In some embodiments, the method includes calculating (such as calculating on a computer), for each locus in the plurality of loci, the expected amount of one or more of the alleles for that locus in the maternal DNA or RNA in the mixed sample using the identity of the allele(s) present at that locus in the obtained phased genetic data of the mother and the fraction of maternal DNA or RNA in

the identity of the allele present at that locus in the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, the number of copies of the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, the identity of one or more possible alleles at that locus in the first or second homologous chromosome segment from the father that is specified by the hypothesis to have been inherited by the fetus, the number of copies of the first or second homologous chromosome segment from the father that is specified by the hypothesis to have been inherited by the fetus, and the fraction of fetal DNA or RNA in the mixed sample. In some embodiments, population frequencies are used to predict the identity of the alleles in the first or second homologous chromosome segment from the father. In some embodiments, the probability for each of the possible alleles at each locus in the first or second homologous chromosome segment from the father are considered to be the same.

**[18]** In some embodiments, the method includes obtaining phased genetic data for both the mother and father of the fetus. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the number of copies of the first homologous chromosome segment or portion thereof from the mother in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion thereof from the mother in the genome of the fetus, the number of copies of the first homologous chromosome segment or portion thereof from the father in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion thereof from the father in the genome of the fetus, and the total number of copies of the chromosome segment of interest present in the genome of the fetus. In some embodiments, the method includes calculating (such as calculating on a computer), for each of the hypotheses, expected genetic data for the plurality of loci in the mixed sample from the obtained phased genetic data from the mother and obtained phased genetic data from the father.

**[19]** In some embodiments, the expected genetic data for each of the hypotheses comprises the identity and an amount of one or more alleles at each locus in the plurality of loci from the maternal DNA or RNA and fetal DNA or RNA in the mixed sample. In some embodiments, the method includes calculating expected genetic data by determining a fraction of fetal DNA or RNA and a fraction of maternal DNA or RNA in the mixed sample. In some embodiments, the method includes calculating (such as calculating on a computer), for each locus in the plurality of loci, the expected amount of one or more of the alleles for that locus in the maternal DNA or RNA in the mixed sample using the identity of the allele(s) present at that locus in the obtained phased genetic data of the mother and the fraction of maternal DNA or RNA in

the mixed sample. In some embodiments, the method includes calculating (such as calculating on a computer), for each locus in the plurality of loci for each hypothesis, the expected amount of one or more of the alleles for that locus in the fetal DNA or RNA in the mixed sample using the identity of the allele present at that locus in the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, the number of copies of the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, the identity of the allele present at that locus in the first or second homologous chromosome segment from the father that is specified by the hypothesis to have been inherited by the fetus, the number of copies of the first or second homologous chromosome segment from the father that is specified by the hypothesis to have been inherited by the fetus, and the fraction of fetal DNA or RNA in the mixed sample.

**[20]** In some embodiments, the method includes calculating (such as calculating on a computer), for each of the hypotheses, a probability distribution of expected genetic data for the plurality of loci in the mixed sample from the obtained phased genetic data from the parent(s). In some embodiments, the method includes increasing the probability in the probability distribution of an a particular allele being present at a first locus in the mixed sample if that particular allele is present in the first homologous segment in the parent and an allele at a nearby locus in the first homologous segment in the parent is observed in the obtained genetic data of the mixed sample; or decreasing the probability in the probability distribution of an a particular allele being present at a first locus in the mixed sample if that particular allele is present in the first homologous segment in the parent and an allele at a nearby locus in the first homologous segment in the parent is not observed in the obtained genetic data of the mixed sample. In some embodiments, the method includes increasing the probability in the probability distribution of an a particular allele being present at a second locus in the mixed sample if that particular allele is present in the second homologous segment in the parent and an allele at a nearby locus in the second homologous segment in the parent is observed in the obtained genetic data of the mixed sample; or decreasing the probability in the probability distribution of an a particular allele being present at a second locus in the mixed sample if that particular allele is present in the second homologous segment in the parent and an allele at a nearby locus in the second homologous segment in the parent is not observed in the obtained genetic data of the mixed sample.

**[21]** In some embodiments, the method includes obtaining phased genetic data for both the mother and father of the fetus. In some embodiments, the method includes enumerating a set of one or more

the mixed sample. In some embodiments, the method includes calculating (such as calculating on a computer), for each locus in the plurality of loci for each hypothesis, the expected amount of one or more of the alleles for that locus in the fetal DNA or RNA in the mixed sample using the identity of the allele present at that locus in the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, the number of copies of the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, the identity of the allele present at that locus in the first or second homologous chromosome segment from the father that is specified by the hypothesis to have been inherited by the fetus, the number of copies of the first or second homologous chromosome segment from the father that is specified by the hypothesis to have been inherited by the fetus, and the fraction of fetal DNA or RNA in the mixed sample.

**[20]** In some embodiments, the method includes calculating (such as calculating on a computer), for each of the hypotheses, a probability distribution of expected genetic data for the plurality of loci in the mixed sample from the obtained phased genetic data from the parent(s). In some embodiments, the method includes increasing the probability in the probability distribution of an a particular allele being present at a first locus in the mixed sample if that particular allele is present in the first homologous segment in the parent and an allele at a nearby locus in the first homologous segment in the parent is observed in the obtained genetic data of the mixed sample; or decreasing the probability in the probability distribution of an a particular allele being present at a first locus in the mixed sample if that particular allele is present in the first homologous segment in the parent and an allele at a nearby locus in the first homologous segment in the parent is not observed in the obtained genetic data of the mixed sample. In some embodiments, the method includes increasing the probability in the probability distribution of an a particular allele being present at a second locus in the mixed sample if that particular allele is present in the second homologous segment in the parent and an allele at a nearby locus in the second homologous segment in the parent is observed in the obtained genetic data of the mixed sample; or decreasing the probability in the probability distribution of an a particular allele being present at a second locus in the mixed sample if that particular allele is present in the second homologous segment in the parent and an allele at a nearby locus in the second homologous segment in the parent is not observed in the obtained genetic data of the mixed sample.

**[21]** In some embodiments, the method includes obtaining phased genetic data for both the mother and father of the fetus. In some embodiments, the method includes enumerating a set of one or more

hypotheses specifying the number of copies of the first homologous chromosome segment or portion thereof from the mother in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion thereof from the mother in the genome of the fetus, the number of copies of the first homologous chromosome segment or portion thereof from the father in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion thereof from the father in the genome of the fetus, and the total number of copies of the chromosome segment of interest present in the genome of the fetus. In some embodiments, the method includes calculating (such as calculating on a computer), for each of the hypotheses, a probability distribution of expected genetic data for the plurality of loci in the mixed sample from the obtained phased genetic data from the mother and father. In some embodiments, the method includes increasing the probability in the probability distribution of an a particular allele being present at a first locus in the mixed sample if that particular allele is present in the first homologous segment in the mother or father and an allele at a nearby locus in the first homologous segment in that parent is observed in the obtained genetic data of the mixed sample; or decreasing the probability in the probability distribution of an a particular allele being present at a first locus in the mixed sample if that particular allele is present in the first homologous segment in the mother or father and an allele at a nearby locus in the first homologous segment in that parent is not observed in the obtained genetic data of the mixed sample. In some embodiments, the method includes increasing the probability in the probability distribution of an a particular allele being present at a second locus in the mixed sample if that particular allele is present in the second homologous segment in the mother or father and an allele at a nearby locus in the second homologous segment in that parent is observed in the obtained genetic data of the mixed sample; or decreasing the probability in the probability distribution of an a particular allele being present at a second locus in the mixed sample if that particular allele is present in the second homologous segment in the mother or father and an allele at a nearby locus in the second homologous segment in that parent is not observed in the obtained genetic data of the mixed sample.

**[22]** In some embodiments, the first locus and the locus that is nearby to the first locus co-segregate. In some embodiments, the second locus and the locus that is nearby to the second locus co-segregate. In some embodiments, no crossovers are expected to occur between the first locus and the locus that is nearby to the first locus. In some embodiments, no crossovers are expected to occur between the second locus and the locus that is nearby to the second locus. In some embodiments, the distance between the first locus and the locus that is nearby to the first locus is less than 5 mb, 1 mb, 100 kb, 10 kb, 1 kb, 0.1 kb, or 0.01 kb. In some embodiments, the distance between the second locus and the locus that is nearby to the second

hypotheses specifying the number of copies of the first homologous chromosome segment or portion thereof from the mother in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion thereof from the mother in the genome of the fetus, the number of copies of the first homologous chromosome segment or portion thereof from the father in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion thereof from the father in the genome of the fetus, and the total number of copies of the chromosome segment of interest present in the genome of the fetus. In some embodiments, the method includes calculating (such as calculating on a computer), for each of the hypotheses, a probability distribution of expected genetic data for the plurality of loci in the mixed sample from the obtained phased genetic data from the mother and father. In some embodiments, the method includes increasing the probability in the probability distribution of an a particular allele being present at a first locus in the mixed sample if that particular allele is present in the first homologous segment in the mother or father and an allele at a nearby locus in the first homologous segment in that parent is observed in the obtained genetic data of the mixed sample; or decreasing the probability in the probability distribution of an a particular allele being present at a first locus in the mixed sample if that particular allele is present in the first homologous segment in the mother or father and an allele at a nearby locus in the first homologous segment in that parent is not observed in the obtained genetic data of the mixed sample. In some embodiments, the method includes increasing the probability in the probability distribution of an a particular allele being present at a second locus in the mixed sample if that particular allele is present in the second homologous segment in the mother or father and an allele at a nearby locus in the second homologous segment in that parent is observed in the obtained genetic data of the mixed sample; or decreasing the probability in the probability distribution of an a particular allele being present at a second locus in the mixed sample if that particular allele is present in the second homologous segment in the mother or father and an allele at a nearby locus in the second homologous segment in that parent is not observed in the obtained genetic data of the mixed sample.

**[22]** In some embodiments, the first locus and the locus that is nearby to the first locus co-segregate. In some embodiments, the second locus and the locus that is nearby to the second locus co-segregate. In some embodiments, no crossovers are expected to occur between the first locus and the locus that is nearby to the first locus. In some embodiments, no crossovers are expected to occur between the second locus and the locus that is nearby to the second locus. In some embodiments, the distance between the first locus and the locus that is nearby to the first locus is less than 5 mb, 1 mb, 100 kb, 10 kb, 1 kb, 0.1 kb, or 0.01 kb. In some embodiments, the distance between the second locus and the locus that is nearby to the second

locus is less than 5 mb, 1 mb, 100 kb, 10 kb, 1 kb, 0.1 kb, or 0.01 kb.

**[23]** In some embodiments, one or more crossovers occurs during the formation of a gamete that contributed a copy of the chromosome segment of interest to the fetus; and the crossover produces a chromosome segment of interest in the genome of the fetus that comprises a portion of the first homologous segment and a portion of the second homologous segment from the parent. In some embodiments, the set of hypothesis comprises one or more hypotheses specifying the number of copies of the chromosome segment of interest in the genome of the fetus that comprises a portion of the first homologous segment and a portion of the second homologous segment from the parent.

**[24]** In some embodiments, the expected genetic data of the mixed sample comprises the expected amount of one or more of the alleles at each locus in the plurality of loci in the mixed sample for each of the hypotheses.

**[25]** In one aspect, the invention features a method of determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of an individual (such as in the genome of one or more cells, cfDNA, cfRNA, an individual suspected of having cancer, a fetus, or an embryo) using phased genetic data. In some embodiments, the method involves simultaneously or sequentially in any order (i) obtaining phased genetic data for the first homologous chromosome segment comprising the identity of the allele present at that locus on the first homologous chromosome segment for each locus in a set of polymorphic loci on the first homologous chromosome segment, (ii) obtaining phased genetic data for the second homologous chromosome segment comprising the identity of the allele present at that locus on the second homologous chromosome segment for each locus in the set of polymorphic loci on the second homologous chromosome segment, and (iii) obtaining measured genetic allelic data comprising the amount of each allele at each of the loci in the set of polymorphic loci in a sample of DNA or RNA from one or more cells from the individual or in a mixed sample of cell- free DNA or RNA from two or more genetically different cells from the individual. In some embodiments, the method involves calculating allele ratios for one or more loci in the set of polymorphic loci that are heterozygous in at least one cell from which the sample was derived. In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus. In some embodiments, the method involves determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment by comparing one or more calculated allele ratios for a locus to an expected allele ratio, such as a ratio that is expected for that locus if the first

locus is less than 5 mb, 1 mb, 100 kb, 10 kb, 1 kb, 0.1 kb, or 0.01 kb.

**[23]** In some embodiments, one or more crossovers occurs during the formation of a gamete that contributed a copy of the chromosome segment of interest to the fetus; and the crossover produces a chromosome segment of interest in the genome of the fetus that comprises a portion of the first homologous segment and a portion of the second homologous segment from the parent. In some embodiments, the set of hypothesis comprises one or more hypotheses specifying the number of copies of the chromosome segment of interest in the genome of the fetus that comprises a portion of the first homologous segment and a portion of the second homologous segment from the parent.

**[24]** In some embodiments, the expected genetic data of the mixed sample comprises the expected amount of one or more of the alleles at each locus in the plurality of loci in the mixed sample for each of the hypotheses.

**[25]** In one aspect, the invention features a method of determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of an individual (such as in the genome of one or more cells, cfDNA, cfRNA, an individual suspected of having cancer, a fetus, or an embryo) using phased genetic data. In some embodiments, the method involves simultaneously or sequentially in any order (i) obtaining phased genetic data for the first homologous chromosome segment comprising the identity of the allele present at that locus on the first homologous chromosome segment for each locus in a set of polymorphic loci on the first homologous chromosome segment, (ii) obtaining phased genetic data for the second homologous chromosome segment comprising the identity of the allele present at that locus on the second homologous chromosome segment for each locus in the set of polymorphic loci on the second homologous chromosome segment, and (iii) obtaining measured genetic allelic data comprising the amount of each allele at each of the loci in the set of polymorphic loci in a sample of DNA or RNA from one or more cells from the individual or in a mixed sample of cell- free DNA or RNA from two or more genetically different cells from the individual. In some embodiments, the method involves calculating allele ratios for one or more loci in the set of polymorphic loci that are heterozygous in at least one cell from which the sample was derived. In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus. In some embodiments, the method involves determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment by comparing one or more calculated allele ratios for a locus to an expected allele ratio, such as a ratio that is expected for that locus if the first

and second homologous chromosome segments are present in equal proportions. In some embodiments, the expected ratio is 0.5 for biallelic loci.

**[26]** In some embodiments for prenatal testing, the method involves simultaneously or sequentially in any order (i) obtaining phased genetic data for the first homologous chromosome segment in the genome of a fetus (such as a fetus gestating in a pregnant mother) comprising the identity of the allele present at that locus on the first homologous chromosome segment for each locus in a set of polymorphic loci on the first homologous chromosome segment, (ii) obtaining phased genetic data for the second homologous chromosome segment in the genome of the fetus comprising the identity of the allele present at that locus on the second homologous chromosome segment for each locus in the set of polymorphic loci on the second homologous chromosome segment, and (iii) obtaining measured genetic allelic data comprising the amount of each allele at each of the loci in the set of polymorphic loci in a mixed sample of DNA or RNA from the mother of the fetus that includes fetal DNA or RNA and maternal DNA or RNA (such as a mixed sample of cell-free DNA or RNA originating from a blood sample from the mother that includes fetal cell-free DNA or RNA and maternal cell-free DNA or RNA). In some embodiments, the method involves calculating allele ratios for one or more loci in the set of polymorphic loci that are heterozygous in the fetus and/or heterozygous in the mother. In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus. In some embodiments, the method involves determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment by comparing one or more calculated allele ratios for a locus to an expected allele ratio, such as a ratio that is expected for that locus if the first and second homologous chromosome segments are present in equal proportions.

**[27]** In some embodiments, a calculated allele ratio is indicative of an overrepresentation of the number of copies of the first homologous chromosome segment if either (i) the allele ratio for the measured quantity of the allele present at that locus on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is greater than the expected allele ratio for that locus, or (ii) the allele ratio for the measured quantity of the allele present at that locus on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than the expected allele ratio for that locus. In some embodiments, a calculated allele ratio is indicative of no overrepresentation of the number of copies of the first homologous chromosome segment if either (i) the allele ratio for the measured quantity of the allele present at that locus on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than or equal to

and second homologous chromosome segments are present in equal proportions. In some embodiments, the expected ratio is 0.5 for biallelic loci.

**[26]** In some embodiments for prenatal testing, the method involves simultaneously or sequentially in any order (i) obtaining phased genetic data for the first homologous chromosome segment in the genome of a fetus (such as a fetus gestating in a pregnant mother) comprising the identity of the allele present at that locus on the first homologous chromosome segment for each locus in a set of polymorphic loci on the first homologous chromosome segment, (ii) obtaining phased genetic data for the second homologous chromosome segment in the genome of the fetus comprising the identity of the allele present at that locus on the second homologous chromosome segment for each locus in the set of polymorphic loci on the second homologous chromosome segment, and (iii) obtaining measured genetic allelic data comprising the amount of each allele at each of the loci in the set of polymorphic loci in a mixed sample of DNA or RNA from the mother of the fetus that includes fetal DNA or RNA and maternal DNA or RNA (such as a mixed sample of cell-free DNA or RNA originating from a blood sample from the mother that includes fetal cell-free DNA or RNA and maternal cell-free DNA or RNA). In some embodiments, the method involves calculating allele ratios for one or more loci in the set of polymorphic loci that are heterozygous in the fetus and/or heterozygous in the mother. In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus. In some embodiments, the method involves determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment by comparing one or more calculated allele ratios for a locus to an expected allele ratio, such as a ratio that is expected for that locus if the first and second homologous chromosome segments are present in equal proportions.

**[27]** In some embodiments, a calculated allele ratio is indicative of an overrepresentation of the number of copies of the first homologous chromosome segment if either (i) the allele ratio for the measured quantity of the allele present at that locus on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is greater than the expected allele ratio for that locus, or (ii) the allele ratio for the measured quantity of the allele present at that locus on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than the expected allele ratio for that locus. In some embodiments, a calculated allele ratio is indicative of no overrepresentation of the number of copies of the first homologous chromosome segment if either (i) the allele ratio for the measured quantity of the allele present at that locus on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than or equal to

the expected allele ratio for that locus, or (ii) the allele ratio for the measured quantity of the allele present at that locus on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is greater than or equal to the expected allele ratio for that locus.

**[28]** In some embodiments, determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment. In some embodiments, predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) are estimated for each hypothesis given the degree of overrepresentation specified by that hypothesis. In some embodiments, the likelihood that the hypothesis is correct is calculated by comparing the calculated allele ratios to the predicted allele ratios, and the hypothesis with the greatest likelihood is selected. In some embodiments, an expected distribution of a test statistic is calculated using the predicted allele ratios for each hypothesis. In some embodiments, the likelihood that the hypothesis is correct is calculated by comparing a test statistic that is calculated using the calculated allele ratios to the expected distribution of the test statistic that is calculated using the predicted allele ratios, and the hypothesis with the greatest likelihood is selected. In some embodiments, predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) are estimated given the phased genetic data for the first homologous chromosome segment, the phased genetic data for the second homologous chromosome segment, and the degree of overrepresentation specified by that hypothesis. In some embodiments, the likelihood that the hypothesis is correct is calculated by comparing the calculated allele ratios to the predicted allele ratios; and the hypothesis with the greatest likelihood is selected.

**[29]** In some embodiments, the ratio of DNA (or RNA) from one or more target cells to the total DNA (or RNA) in the sample is calculated. An exemplary ratio is the ratio of fetal DNA (or RNA) to the total DNA (or RNA) in the sample. In some embodiments, the ratio of fetal DNA to total DNA in the sample is determined by measuring the amount of an allele at one or more loci in which the fetus has the allele and the mother does not have the allele. In some embodiments, the ratio of fetal DNA to total DNA in the sample is determined by measuring the difference in methylation between one or more maternal and fetal alleles. In some embodiments, a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment are enumerated. In some embodiments, predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) are estimated given the calculated ratio of DNA or RNA and the

the expected allele ratio for that locus, or (ii) the allele ratio for the measured quantity of the allele present at that locus on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is greater than or equal to the expected allele ratio for that locus.

**[28]** In some embodiments, determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment. In some embodiments, predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) are estimated for each hypothesis given the degree of overrepresentation specified by that hypothesis. In some embodiments, the likelihood that the hypothesis is correct is calculated by comparing the calculated allele ratios to the predicted allele ratios, and the hypothesis with the greatest likelihood is selected. In some embodiments, an expected distribution of a test statistic is calculated using the predicted allele ratios for each hypothesis. In some embodiments, the likelihood that the hypothesis is correct is calculated by comparing a test statistic that is calculated using the calculated allele ratios to the expected distribution of the test statistic that is calculated using the predicted allele ratios, and the hypothesis with the greatest likelihood is selected. In some embodiments, predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) are estimated given the phased genetic data for the first homologous chromosome segment, the phased genetic data for the second homologous chromosome segment, and the degree of overrepresentation specified by that hypothesis. In some embodiments, the likelihood that the hypothesis is correct is calculated by comparing the calculated allele ratios to the predicted allele ratios; and the hypothesis with the greatest likelihood is selected.

**[29]** In some embodiments, the ratio of DNA (or RNA) from one or more target cells to the total DNA (or RNA) in the sample is calculated. An exemplary ratio is the ratio of fetal DNA (or RNA) to the total DNA (or RNA) in the sample. In some embodiments, the ratio of fetal DNA to total DNA in the sample is determined by measuring the amount of an allele at one or more loci in which the fetus has the allele and the mother does not have the allele. In some embodiments, the ratio of fetal DNA to total DNA in the sample is determined by measuring the difference in methylation between one or more maternal and fetal alleles. In some embodiments, a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment are enumerated. In some embodiments, predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) are estimated given the calculated ratio of DNA or RNA and the

degree of overrepresentation specified by that hypothesis are estimated for each hypothesis. In some embodiments, the likelihood that the hypothesis is correct is calculated by comparing the calculated allele ratios to the predicted allele ratios, and the hypothesis with the greatest likelihood is selected. In some embodiments, an expected distribution of a test statistic calculated using the predicted allele ratios and the calculated ratio of DNA or RNA is estimated for each hypothesis. In some embodiments, the likelihood that the hypothesis is correct is determined by comparing a test statistic calculated using the calculated allele ratios and the calculated ratio of DNA or RNA to the expected distribution of the test statistic calculated using the predicted allele ratios and the calculated ratio of DNA or RNA, and the hypothesis with the greatest likelihood is selected.

**[30]** In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment. In some embodiments, the method includes estimating, for each hypothesis, either (i) predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) given the degree of overrepresentation specified by that hypothesis or (ii) for one or more possible ratios of DNA or RNA (such as ratios of fetal DNA or RNA to the total DNA or RNA in the sample), an expected distribution of a test statistic calculated using the predicted allele ratios and the possible ratio of DNA or RNA from the one or more target cells (such as fetal cells) to the total DNA or RNA in the sample. In some embodiments, a data fit is calculated by comparing either (i) the calculated allele ratios to the predicted allele ratios, or (ii) a test statistic calculated using the calculated allele ratios and the possible ratio of DNA or RNA to the expected distribution of the test statistic calculated using the predicted allele ratios and the possible ratio of DNA or RNA. In some embodiments, one or more of the hypotheses are ranked according to the data fit, and the hypothesis that is ranked the highest is selected. In some embodiments, a technique or algorithm, such as a search algorithm, is used for one or more of the following steps: calculating the data fit, ranking the hypotheses, or selecting the hypothesis that is ranked the highest. In some embodiments, the data fit is a fit to a beta-binomial distribution or a fit to a binomial distribution. In some embodiments, the technique or algorithm is selected from the group consisting of maximum likelihood estimation, maximum a-posteriori estimation, Bayesian estimation, dynamic estimation (such as dynamic Bayesian estimation), and expectation- maximization estimation. In some embodiments, the method includes applying the technique or algorithm to the obtained genetic data and the expected genetic data.

**[31]** In some embodiments, the method includes creating a partition of possible ratios (such as ratios

degree of overrepresentation specified by that hypothesis are estimated for each hypothesis. In some embodiments, the likelihood that the hypothesis is correct is calculated by comparing the calculated allele ratios to the predicted allele ratios, and the hypothesis with the greatest likelihood is selected. In some embodiments, an expected distribution of a test statistic calculated using the predicted allele ratios and the calculated ratio of DNA or RNA is estimated for each hypothesis. In some embodiments, the likelihood that the hypothesis is correct is determined by comparing a test statistic calculated using the calculated allele ratios and the calculated ratio of DNA or RNA to the expected distribution of the test statistic calculated using the predicted allele ratios and the calculated ratio of DNA or RNA, and the hypothesis with the greatest likelihood is selected.

**[30]** In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment. In some embodiments, the method includes estimating, for each hypothesis, either (i) predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) given the degree of overrepresentation specified by that hypothesis or (ii) for one or more possible ratios of DNA or RNA (such as ratios of fetal DNA or RNA to the total DNA or RNA in the sample), an expected distribution of a test statistic calculated using the predicted allele ratios and the possible ratio of DNA or RNA from the one or more target cells (such as fetal cells) to the total DNA or RNA in the sample. In some embodiments, a data fit is calculated by comparing either (i) the calculated allele ratios to the predicted allele ratios, or (ii) a test statistic calculated using the calculated allele ratios and the possible ratio of DNA or RNA to the expected distribution of the test statistic calculated using the predicted allele ratios and the possible ratio of DNA or RNA. In some embodiments, one or more of the hypotheses are ranked according to the data fit, and the hypothesis that is ranked the highest is selected. In some embodiments, a technique or algorithm, such as a search algorithm, is used for one or more of the following steps: calculating the data fit, ranking the hypotheses, or selecting the hypothesis that is ranked the highest. In some embodiments, the data fit is a fit to a beta-binomial distribution or a fit to a binomial distribution. In some embodiments, the technique or algorithm is selected from the group consisting of maximum likelihood estimation, maximum a-posteriori estimation, Bayesian estimation, dynamic estimation (such as dynamic Bayesian estimation), and expectation- maximization estimation. In some embodiments, the method includes applying the technique or algorithm to the obtained genetic data and the expected genetic data.

**[31]** In some embodiments, the method includes creating a partition of possible ratios (such as ratios

of fetal DNA or RNA to the total DNA or RNA in the sample) that range from a lower limit to an upper limit for the ratio of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample. In some embodiments, a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment are enumerated. In some embodiments, the method includes estimating, for each of the possible ratios of DNA or RNA in the partition and for each hypothesis, either (i) predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) given the possible ratio of DNA or RNA and the degree of overrepresentation specified by that hypothesis or (ii) an expected distribution of a test statistic calculated using the predicted allele ratios and the possible ratio of DNA or RNA. In some embodiments, the method includes calculating, for each of the possible ratios of DNA or RNA in the partition and for each hypothesis, the likelihood that the hypothesis is correct by comparing either (i) the calculated allele ratios to the predicted allele ratios, or (ii) a test statistic calculated using the calculated allele ratios and the possible ratio of DNA or RNA to the expected distribution of the test statistic calculated using the predicted allele ratios and the possible ratio of DNA or RNA. In some embodiments, the combined probability for each hypothesis is determined by combining the probabilities of that hypothesis for each of the possible ratios in the partition; and the hypothesis with the greatest combined probability is selected. In some embodiments, the combined probability for each hypothesis is determined by weighting the probability of a hypothesis for a particular possible ratio based on the likelihood that the possible ratio is the correct ratio.

**[32]** In one aspect, the invention features a method for determining a number of copies of a chromosome or chromosome segment in the genome of one or more cells from an individual using phased or unphased genetic data. In some embodiments, the method involves obtaining genetic data at a set of polymorphic loci on the chromosome or chromosome segment in a sample by measuring the quantity of each allele at each locus. In some embodiments, the sample is a sample of DNA or RNA from one or more cells from the individual or a mixed sample of cell-free DNA from the individual that includes cell-free DNA from two or more genetically different cells. In some embodiments, allele ratios are calculated for the loci that are heterozygous in at least one cell from which the sample was derived. In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus. In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles (such as the allele on the first homologous chromosome segment) divided by the measured quantity of one or more other alleles

of fetal DNA or RNA to the total DNA or RNA in the sample) that range from a lower limit to an upper limit for the ratio of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample. In some embodiments, a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment are enumerated. In some embodiments, the method includes estimating, for each of the possible ratios of DNA or RNA in the partition and for each hypothesis, either (i) predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) given the possible ratio of DNA or RNA and the degree of overrepresentation specified by that hypothesis or (ii) an expected distribution of a test statistic calculated using the predicted allele ratios and the possible ratio of DNA or RNA. In some embodiments, the method includes calculating, for each of the possible ratios of DNA or RNA in the partition and for each hypothesis, the likelihood that the hypothesis is correct by comparing either (i) the calculated allele ratios to the predicted allele ratios, or (ii) a test statistic calculated using the calculated allele ratios and the possible ratio of DNA or RNA to the expected distribution of the test statistic calculated using the predicted allele ratios and the possible ratio of DNA or RNA. In some embodiments, the combined probability for each hypothesis is determined by combining the probabilities of that hypothesis for each of the possible ratios in the partition; and the hypothesis with the greatest combined probability is selected. In some embodiments, the combined probability for each hypothesis is determined by weighting the probability of a hypothesis for a particular possible ratio based on the likelihood that the possible ratio is the correct ratio.

**[32]** In one aspect, the invention features a method for determining a number of copies of a chromosome or chromosome segment in the genome of one or more cells from an individual using phased or unphased genetic data. In some embodiments, the method involves obtaining genetic data at a set of polymorphic loci on the chromosome or chromosome segment in a sample by measuring the quantity of each allele at each locus. In some embodiments, the sample is a sample of DNA or RNA from one or more cells from the individual or a mixed sample of cell-free DNA from the individual that includes cell-free DNA from two or more genetically different cells. In some embodiments, allele ratios are calculated for the loci that are heterozygous in at least one cell from which the sample was derived. In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus. In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles (such as the allele on the first homologous chromosome segment) divided by the measured quantity of one or more other alleles

(such as the allele on the second homologous chromosome segment) for the locus. In some embodiments, a set of one or more hypotheses specifying the number of copies of the chromosome or chromosome segment in the genome of one or more of the cells are enumerated. In some embodiments, the hypothesis that is most likely based on the test statistic is selected, thereby determining the number of copies of the chromosome or chromosome segment in the genome of one or more of the cells.

**[33]** In one aspect, the invention features a method for determining a number of copies of a chromosome or chromosome segment in the genome of a fetus (such as a fetus that is gestating in a pregnant mother) using phased or unphased genetic data. In some embodiments, the method involves obtaining genetic data at a set of polymorphic loci on the chromosome or chromosome segment in a sample by measuring the quantity of each allele at each locus. In some embodiments, the sample is a mixed sample of DNA comprising fetal DNA or RNA and maternal DNA or RNA from the mother of the fetus (such as a mixed sample of cell-free DNA or RNA originating from a blood sample from the mother that includes fetal cell-free DNA or

RNA and maternal cell-free DNA or RNA). In some embodiments, allele ratios are calculated for the loci that are heterozygous in the fetus and/or heterozygous in the mother. In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus. In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles (such as the allele on the first homologous chromosome segment) divided by the measured quantity of one or more other alleles (such as the allele on the second homologous chromosome segment) for the locus. In some embodiments, a set of one or more hypotheses specifying the number of copies of the chromosome or chromosome segment in the genome of fetus are enumerated. In some embodiments, the hypothesis that is most likely based on the test statistic is selected, thereby determining the number of copies of the chromosome or chromosome segment in the genome of the fetus.

**[34]** In some embodiments, a hypothesis is selected if the probability that the test statistic belongs to a distribution of the test statistic for that hypothesis is above an upper threshold; one or more of the hypotheses is rejected if the probability that the test statistic belongs to the distribution of the test statistic for that hypothesis is below a lower threshold; or a hypothesis is neither selected nor rejected if the probability that the test statistic belongs to the distribution of the test statistic for that hypothesis is between the lower threshold and the upper threshold, or if the probability is not determined with sufficiently high confidence. In some embodiments, the overrepresentation of the number of copies of the first homologous

(such as the allele on the second homologous chromosome segment) for the locus. In some embodiments, a set of one or more hypotheses specifying the number of copies of the chromosome or chromosome segment in the genome of one or more of the cells are enumerated. In some embodiments, the hypothesis that is most likely based on the test statistic is selected, thereby determining the number of copies of the chromosome or chromosome segment in the genome of one or more of the cells.

**[33]** In one aspect, the invention features a method for determining a number of copies of a chromosome or chromosome segment in the genome of a fetus (such as a fetus that is gestating in a pregnant mother) using phased or unphased genetic data. In some embodiments, the method involves obtaining genetic data at a set of polymorphic loci on the chromosome or chromosome segment in a sample by measuring the quantity of each allele at each locus. In some embodiments, the sample is a mixed sample of DNA comprising fetal DNA or RNA and maternal DNA or RNA from the mother of the fetus (such as a mixed sample of cell-free DNA or RNA originating from a blood sample from the mother that includes fetal cell-free DNA or

RNA and maternal cell-free DNA or RNA). In some embodiments, allele ratios are calculated for the loci that are heterozygous in the fetus and/or heterozygous in the mother. In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus. In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles (such as the allele on the first homologous chromosome segment) divided by the measured quantity of one or more other alleles (such as the allele on the second homologous chromosome segment) for the locus. In some embodiments, a set of one or more hypotheses specifying the number of copies of the chromosome or chromosome segment in the genome of fetus are enumerated. In some embodiments, the hypothesis that is most likely based on the test statistic is selected, thereby determining the number of copies of the chromosome or chromosome segment in the genome of the fetus.

**[34]** In some embodiments, a hypothesis is selected if the probability that the test statistic belongs to a distribution of the test statistic for that hypothesis is above an upper threshold; one or more of the hypotheses is rejected if the probability that the test statistic belongs to the distribution of the test statistic for that hypothesis is below a lower threshold; or a hypothesis is neither selected nor rejected if the probability that the test statistic belongs to the distribution of the test statistic for that hypothesis is between the lower threshold and the upper threshold, or if the probability is not determined with sufficiently high confidence. In some embodiments, the overrepresentation of the number of copies of the first homologous

chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment. In some embodiments, the total measured quantity of all the alleles for one or more of the loci is compared to a reference amount to determine whether the overrepresentation of the number of copies of the first homologous chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment. In some embodiments, the magnitude of the difference between the calculated allele ratio and the expected allele ratio for one or more loci is used to determine whether the overrepresentation of the number of copies of the first homologous chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment. In some embodiments, the first and second homologous chromosome segments are determined to be present in equal proportions if there is not an overrepresentation of the number of copies of the first homologous chromosome segment, and there is not an overrepresentation of the second homologous chromosome segment (such as in the genome of the cells, cfDNA, cfRNA, individual, fetus, or embryo).

**[35]** In some embodiments, the ratio of DNA from the one or more target cells to the total DNA in the sample is determined based on the total or relative amount of one or more alleles at one or more loci for which the genotype of the target cells differs from the genotype of the nontarget cells and for which the target cells and non-target cells are expected to be disomic. In some embodiments, this ratio is used to determine whether the overrepresentation of the number of copies of the first homologous chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment. In some embodiments, the ratio is used to determine the number of extra copies of a chromosome segment or chromosome that is duplicated. In some embodiments, the phased genetic data includes probabilistic data. In some embodiments, obtaining the phased genetic data for the first homologous chromosome segment and/or the second homologous chromosome segment in the genome of the fetus includes obtaining phased genetic data for the first homologous chromosome segment and/or the second homologous chromosome segment in the genome of one or both biological parents of the fetus, and inferring which homologous chromosome segment the fetus inherited from one or both biological parents. In some embodiments, the probability of one or more crossovers (such as 1, 2, 3, or 4 crossovers) that may have occurred during the formation of a gamete that contributed a copy of the first homologous chromosome segment or the second homologous chromosome segment to the fetus individual is used to infer which homologous chromosome segment(s) the fetus

chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment. In some embodiments, the total measured quantity of all the alleles for one or more of the loci is compared to a reference amount to determine whether the overrepresentation of the number of copies of the first homologous chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment. In some embodiments, the magnitude of the difference between the calculated allele ratio and the expected allele ratio for one or more loci is used to determine whether the overrepresentation of the number of copies of the first homologous chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment. In some embodiments, the first and second homologous chromosome segments are determined to be present in equal proportions if there is not an overrepresentation of the number of copies of the first homologous chromosome segment, and there is not an overrepresentation of the second homologous chromosome segment (such as in the genome of the cells, cfDNA, cfRNA, individual, fetus, or embryo).

**[35]** In some embodiments, the ratio of DNA from the one or more target cells to the total DNA in the sample is determined based on the total or relative amount of one or more alleles at one or more loci for which the genotype of the target cells differs from the genotype of the nontarget cells and for which the target cells and non-target cells are expected to be disomic. In some embodiments, this ratio is used to determine whether the overrepresentation of the number of copies of the first homologous chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment. In some embodiments, the ratio is used to determine the number of extra copies of a chromosome segment or chromosome that is duplicated. In some embodiments, the phased genetic data includes probabilistic data. In some embodiments, obtaining the phased genetic data for the first homologous chromosome segment and/or the second homologous chromosome segment in the genome of the fetus includes obtaining phased genetic data for the first homologous chromosome segment and/or the second homologous chromosome segment in the genome of one or both biological parents of the fetus, and inferring which homologous chromosome segment the fetus inherited from one or both biological parents. In some embodiments, the probability of one or more crossovers (such as 1, 2, 3, or 4 crossovers) that may have occurred during the formation of a gamete that contributed a copy of the first homologous chromosome segment or the second homologous chromosome segment to the fetus individual is used to infer which homologous chromosome segment(s) the fetus

inherited from one or both biological parents. In some embodiments, phased genetic data for the mother and/or father of the fetus is obtained using a technique selected from the group consisting of digital PCR, inferring a haplotype using population based haplotype frequencies, haplotyping using a haploid cell such as a sperm or egg, haplotyping using genetic data from one or more first degree relatives, and combinations thereof. In some embodiments, the phased genetic data for the individual is obtained by phasing a portion or all of region corresponding to a deletion or duplication in a sample from the individual. In some embodiments, the phased genetic data for a fetus is obtained by phasing a portion or all of region corresponding to a deletion or duplication in a sample from the fetus or the mother of the fetus. In some embodiments, obtaining phased genetic data for the first and second homologous chromosome segments includes determining the identity of alleles present in one of the chromosome segments and determining the identity of alleles present in the other chromosome segment by inference. In some embodiments, alleles from unphased genetic data that are not present in the first homologous chromosome segment are assigned to the second homologous chromosome segment. For example, if the genotype of the individual is (AB, AB) and the phased data for the individual indicates that the first haplotype is (A, A); then, the other haplotype can be inferred to be (B, B). In some embodiments, if only one allele is measured at a locus then that allele is determined to be part of both the first and second homologous chromosome segments (e.g., if the genotype is AA at a locus than both haplotypes have the A allele). In some embodiments, the phased genetic data for the individual comprises determining whether or not one or more possible chromosome crossovers occurred, such as by determining the sequence of a recombination hotspot and optionally of a region flanking a recombination hotspot. In some embodiments, any of the primer libraries of the invention are used to detect a recombination event to determine what haplotype blocks are present in the genome of an individual.

**[36]** In some embodiments, the method includes using a joint distribution model (such as a joint distribution model that takes into account the linkage between loci), performing a linkage analysis, using a binomial distribution model, using a beta-binomial distribution model, and/or using the likelihood of crossovers having occurred during the meiosis that gave rise to the gametes that formed the embryo that grew into the fetus (such as using the probability of chromosomes crossing over at different locations in a chromosome to model dependence between polymorphic alleles on the chromosome or chromosome segment of interest).

**[37]** In some embodiments, one or more of the calculated allele ratios for the cfDNA or cfRNA are indicative of the corresponding allele ratios for DNA or RNA in the cells from which the cfDNA or

inherited from one or both biological parents. In some embodiments, phased genetic data for the mother and/or father of the fetus is obtained using a technique selected from the group consisting of digital PCR, inferring a haplotype using population based haplotype frequencies, haplotyping using a haploid cell such as a sperm or egg, haplotyping using genetic data from one or more first degree relatives, and combinations thereof. In some embodiments, the phased genetic data for the individual is obtained by phasing a portion or all of region corresponding to a deletion or duplication in a sample from the individual. In some embodiments, the phased genetic data for a fetus is obtained by phasing a portion or all of region corresponding to a deletion or duplication in a sample from the fetus or the mother of the fetus. In some embodiments, obtaining phased genetic data for the first and second homologous chromosome segments includes determining the identity of alleles present in one of the chromosome segments and determining the identity of alleles present in the other chromosome segment by inference. In some embodiments, alleles from unphased genetic data that are not present in the first homologous chromosome segment are assigned to the second homologous chromosome segment. For example, if the genotype of the individual is (AB, AB) and the phased data for the individual indicates that the first haplotype is (A, A); then, the other haplotype can be inferred to be (B, B). In some embodiments, if only one allele is measured at a locus then that allele is determined to be part of both the first and second homologous chromosome segments (e.g., if the genotype is AA at a locus than both haplotypes have the A allele). In some embodiments, the phased genetic data for the individual comprises determining whether or not one or more possible chromosome crossovers occurred, such as by determining the sequence of a recombination hotspot and optionally of a region flanking a recombination hotspot. In some embodiments, any of the primer libraries of the invention are used to detect a recombination event to determine what haplotype blocks are present in the genome of an individual.

**[36]** In some embodiments, the method includes using a joint distribution model (such as a joint distribution model that takes into account the linkage between loci), performing a linkage analysis, using a binomial distribution model, using a beta-binomial distribution model, and/or using the likelihood of crossovers having occurred during the meiosis that gave rise to the gametes that formed the embryo that grew into the fetus (such as using the probability of chromosomes crossing over at different locations in a chromosome to model dependence between polymorphic alleles on the chromosome or chromosome segment of interest).

**[37]** In some embodiments, one or more of the calculated allele ratios for the cfDNA or cfRNA are indicative of the corresponding allele ratios for DNA or RNA in the cells from which the cfDNA or

cfRNA was derived. In some embodiments, one or more of the calculated allele ratios for the cfDNA or cfRNA are indicative of the corresponding allele ratios in the genome of the individual. In some embodiments, an allele ratio is only calculated or is only compared to an expected allele ratio if the measured genetic data indicate that more than one different allele is present for that locus in the sample (such as in a cfDNA or cfRNA sample). In some embodiments, an allele ratio is only calculated or is only compared to an expected allele ratio if the locus is heterozygous in at least one of the cells from which the sample was derived (such as a locus that is heterozygous in the fetus and/or heterozygous in the mother). In some embodiments, an allele ratio is only calculated or is only compared to an expected allele ratio if the locus is heterozygous in the fetus. In some embodiments, an allele ratio is calculated and compared to an expected allele ratio for a homozygous locus. For example, allele ratios for loci that are predicted to be homozygous for a particular individual being tested (or for both a fetus and pregnant mother) may be analyzed to determine the level of noise or error in the system.

**[38]** In some embodiments, at least 10; 50; 100; 200; 300; 500; 750; 1,000; 2,000; 3,000; 4,000, or more loci (such as SNPs) are analyzed for a chromosome or chromosome segment of interest. In some embodiments, the average number of loci (such as SNPs) per mb in a chromosome or chromosome segment of interest is at least 1; 10; 25; 50; 100; 150; 200; 300; 500; 750; 1,000; or more loci per mb. In some embodiments, the average number of loci (such as SNPs) per mb in a chromosome or chromosome segment of interest is between 1 and 500 loci per mb, such as between 1 and 50, 50 and 100, 100 and 200, 200 and 400, 200 and 300, or 300 and 400 loci per mb, inclusive. In some embodiments, loci in multiple portions of a potential deletion or duplication are analyzed to increase the sensitivity and/or specificity of the CNV determination compared to only analyzing 1 loci or only analyzing a few loci that are near each other. In some embodiments, only the two most common alleles at each locus are measured or are used to determine the calculated allele ratio. In some embodiments, the amplification of loci is performed using a polymerase (*e.g.*, a DNA polymerase, RNA polymerase, or reverse transcriptase) with low 5'→3' exonuclease and/or low strand displacement activity. In some embodiments, the measured genetic allelic data is obtained by (i) sequencing the DNA or RNA in the sample, (ii) amplifying DNA or RNA in the sample and then sequencing the amplified DNA, or (ii) amplifying the DNA or RNA in the sample, ligating PCR products, and then sequencing the ligated products. In some embodiments, measured genetic allelic data is obtained by dividing the DNA or RNA from the sample into a plurality of fractions, adding a different barcode to the DNA or RNA in each fraction (*e.g.*, such that all the DNA or RNA in a particular fraction has the same barcode), optionally amplifying the barcoded DNA or RNA, combining the

cfRNA was derived. In some embodiments, one or more of the calculated allele ratios for the cfDNA or cfRNA are indicative of the corresponding allele ratios in the genome of the individual. In some embodiments, an allele ratio is only calculated or is only compared to an expected allele ratio if the measured genetic data indicate that more than one different allele is present for that locus in the sample (such as in a cfDNA or cfRNA sample). In some embodiments, an allele ratio is only calculated or is only compared to an expected allele ratio if the locus is heterozygous in at least one of the cells from which the sample was derived (such as a locus that is heterozygous in the fetus and/or heterozygous in the mother). In some embodiments, an allele ratio is only calculated or is only compared to an expected allele ratio if the locus is heterozygous in the fetus. In some embodiments, an allele ratio is calculated and compared to an expected allele ratio for a homozygous locus. For example, allele ratios for loci that are predicted to be homozygous for a particular individual being tested (or for both a fetus and pregnant mother) may be analyzed to determine the level of noise or error in the system.

**[38]** In some embodiments, at least 10; 50; 100; 200; 300; 500; 750; 1,000; 2,000; 3,000; 4,000, or more loci (such as SNPs) are analyzed for a chromosome or chromosome segment of interest. In some embodiments, the average number of loci (such as SNPs) per mb in a chromosome or chromosome segment of interest is at least 1; 10; 25; 50; 100; 150; 200; 300; 500; 750; 1,000; or more loci per mb. In some embodiments, the average number of loci (such as SNPs) per mb in a chromosome or chromosome segment of interest is between 1 and 500 loci per mb, such as between 1 and 50, 50 and 100, 100 and 200, 200 and 400, 200 and 300, or 300 and 400 loci per mb, inclusive. In some embodiments, loci in multiple portions of a potential deletion or duplication are analyzed to increase the sensitivity and/or specificity of the CNV determination compared to only analyzing 1 loci or only analyzing a few loci that are near each other. In some embodiments, only the two most common alleles at each locus are measured or are used to determine the calculated allele ratio. In some embodiments, the amplification of loci is performed using a polymerase (*e.g.*, a DNA polymerase, RNA polymerase, or reverse transcriptase) with low 5'→3' exonuclease and/or low strand displacement activity. In some embodiments, the measured genetic allelic data is obtained by (i) sequencing the DNA or RNA in the sample, (ii) amplifying DNA or RNA in the sample and then sequencing the amplified DNA, or (ii) amplifying the DNA or RNA in the sample, ligating PCR products, and then sequencing the ligated products. In some embodiments, measured genetic allelic data is obtained by dividing the DNA or RNA from the sample into a plurality of fractions, adding a different barcode to the DNA or RNA in each fraction (*e.g.*, such that all the DNA or RNA in a particular fraction has the same barcode), optionally amplifying the barcoded DNA or RNA, combining the

fractions, and then sequencing the barcoded DNA or RNA in the combined fractions. In some embodiments, alleles of the polymorphic loci (such as SNPs) are identified using one or more of the following methods: sequencing (such as nanopore sequencing or Halcyon Molecular sequencing), SNP array, real time PCR, TaqMan, Nanostring nCounter® Analysis System, Illumina GoldenGate Genotyping Assay that uses a discriminatory DNA polymerase and ligase, ligation-mediated PCR, or Linked Inverted Probes (LIPs; which can also be called precircularized probes, pre-circularizing probes, circularizing probes, Padlock Probes, or Molecular Inversion Probes (MIPs)). In some embodiments, two or more (such as 3 or 4) target amplicons are ligated together and then the ligated products are sequenced. In some embodiments, measurements for different alleles for the same locus are adjusted for differences in metabolism, apoptosis, histones, inactivation, and/or amplification between the alleles (such as differences in amplification efficiency between different alleles of the same locus). In some embodiments, this adjustment is performed prior to calculating allele ratios for the obtained genetic data or prior to comparing the measured genetic data to the expected genetic data.

**[39]** In some embodiments, the method also includes determining the presence or absence of one or more risk factors for a disease or disorder. In some embodiments, the method also includes determining the presence or absence of one or more polymorphisms or mutations associated with the disease or disorder or an increased risk for a disease or disorder. In some embodiments, the method also includes determining the total level of cfDNA cf mRNA, cf nDNA, cfRNA, miRNA, or any combination thereof. In some embodiments, the method includes determining the level of one or more cfDNA cf mRNA, cf nDNA, cfRNA, and/or miRNA molecules of interest, such as molecules with a polymorphism or mutation associated with a disease or disorder or an increased risk for a disease or disorder. In some embodiments, the fraction of tumor DNA out of total DNA (such as the fraction of tumor cfDNA out of total cfDNA or the fraction of tumor cfDNA with a particular mutation out of total cfDNA) is determined. In some embodiments, this tumor fraction is used to determine the stage of a cancer (since higher tumor fractions can be associated with more advanced stages of cancer). In some embodiments, the method also includes determining the total level of DNA or RNA level. In some embodiments, the method includes determining the methylation level of one or more DNA or RNA molecules of interest, such as molecules with a polymorphism or mutation associated with a disease or disorder or an increased risk for a disease or disorder. In some embodiments, the method includes determining the presence or absence of a change in DNA integrity. In some embodiments, the method also includes determining the total level of mRNA splicing. In some embodiments, the method includes determining the level of mRNA splicing or detecting

fractions, and then sequencing the barcoded DNA or RNA in the combined fractions. In some embodiments, alleles of the polymorphic loci (such as SNPs) are identified using one or more of the following methods: sequencing (such as nanopore sequencing or Halcyon Molecular sequencing), SNP array, real time PCR, TaqMan, Nanostring nCounter® Analysis System, Illumina GoldenGate Genotyping Assay that uses a discriminatory DNA polymerase and ligase, ligation-mediated PCR, or Linked Inverted Probes (LIPs; which can also be called precircularized probes, pre-circularizing probes, circularizing probes, Padlock Probes, or Molecular Inversion Probes (MIPs)). In some embodiments, two or more (such as 3 or 4) target amplicons are ligated together and then the ligated products are sequenced. In some embodiments, measurements for different alleles for the same locus are adjusted for differences in metabolism, apoptosis, histones, inactivation, and/or amplification between the alleles (such as differences in amplification efficiency between different alleles of the same locus). In some embodiments, this adjustment is performed prior to calculating allele ratios for the obtained genetic data or prior to comparing the measured genetic data to the expected genetic data.

**[39]** In some embodiments, the method also includes determining the presence or absence of one or more risk factors for a disease or disorder. In some embodiments, the method also includes determining the presence or absence of one or more polymorphisms or mutations associated with the disease or disorder or an increased risk for a disease or disorder. In some embodiments, the method also includes determining the total level of cfDNA cf mRNA, cf nDNA, cfRNA, miRNA, or any combination thereof. In some embodiments, the method includes determining the level of one or more cfDNA cf mRNA, cf nDNA, cfRNA, and/or miRNA molecules of interest, such as molecules with a polymorphism or mutation associated with a disease or disorder or an increased risk for a disease or disorder. In some embodiments, the fraction of tumor DNA out of total DNA (such as the fraction of tumor cfDNA out of total cfDNA or the fraction of tumor cfDNA with a particular mutation out of total cfDNA) is determined. In some embodiments, this tumor fraction is used to determine the stage of a cancer (since higher tumor fractions can be associated with more advanced stages of cancer). In some embodiments, the method also includes determining the total level of DNA or RNA level. In some embodiments, the method includes determining the methylation level of one or more DNA or RNA molecules of interest, such as molecules with a polymorphism or mutation associated with a disease or disorder or an increased risk for a disease or disorder. In some embodiments, the method includes determining the presence or absence of a change in DNA integrity. In some embodiments, the method also includes determining the total level of mRNA splicing. In some embodiments, the method includes determining the level of mRNA splicing or detecting

alternative mRNA splicing for one or RNA molecules of interest, such as molecules with a polymorphism or mutation associated with a disease or disorder or an increased risk for a disease or disorder.

**[40]** In some embodiments, the invention features a method for detecting a cancer phenotype in an individual, wherein the cancer phenotype is defined by the presence of at least one of a set of mutations. In some embodiments, the method includes obtaining DNA or RNA measurements for a sample of DNA or RNA from one or more cells from the individual, wherein one or more of the cells is suspected of having the cancer phenotype; and analyzing the DNA or RNA measurements to determine, for each of the mutations in the set of mutations, the likelihood that at least one of the cells has that mutation. In some embodiments, the method includes determining that the individual has the cancer phenotype if either (i) for at least one of the mutations, the likelihood that at least one of the cells contains that mutations is greater than a threshold, or (ii) for at least one of the mutations, the likelihood that at least one of the cells has that mutations is less than the threshold, and for a plurality of the mutations, the combined likelihood that at least one of the cells has at least one of the mutations is greater than the threshold. In some embodiments, one or more cells have a subset or all of the mutations in the set of mutations. In some embodiments, the subset of mutations is associated with cancer or an increased risk for cancer. In some embodiments, the sample includes cell-free DNA or RNA. In some embodiments, the DNA or RNA measurements include measurements (such as the quantity of each allele at each locus) at a set of polymorphic loci on one or more chromosomes or chromosome segments of interest.

**[41]** In one aspect, the invention features methods for selecting a therapy for the treatment, stabilization, or prevention of a disease or disorder in a mammal. In some embodiments, the method includes determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment using any of the methods described herein. In some embodiments, a therapy is selected for the mammal (such as a therapy for a disease or disorder associated with the overrepresentation of the first homologous chromosome segment).

**[42]** In one aspect, the invention features methods for preventing, delaying, stabilizing, or treating a disease or disorder in a mammal. In some embodiments, the method includes determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment using any of the methods described herein. In some embodiments, a therapy is selected for the mammal (such as a therapy for a disease or disorder associated with the overrepresentation of the first homologous chromosome segment) and then the therapy is

alternative mRNA splicing for one or RNA molecules of interest, such as molecules with a polymorphism or mutation associated with a disease or disorder or an increased risk for a disease or disorder.

**[40]** In some embodiments, the invention features a method for detecting a cancer phenotype in an individual, wherein the cancer phenotype is defined by the presence of at least one of a set of mutations. In some embodiments, the method includes obtaining DNA or RNA measurements for a sample of DNA or RNA from one or more cells from the individual, wherein one or more of the cells is suspected of having the cancer phenotype; and analyzing the DNA or RNA measurements to determine, for each of the mutations in the set of mutations, the likelihood that at least one of the cells has that mutation. In some embodiments, the method includes determining that the individual has the cancer phenotype if either (i) for at least one of the mutations, the likelihood that at least one of the cells contains that mutations is greater than a threshold, or (ii) for at least one of the mutations, the likelihood that at least one of the cells has that mutations is less than the threshold, and for a plurality of the mutations, the combined likelihood that at least one of the cells has at least one of the mutations is greater than the threshold. In some embodiments, one or more cells have a subset or all of the mutations in the set of mutations. In some embodiments, the subset of mutations is associated with cancer or an increased risk for cancer. In some embodiments, the sample includes cell-free DNA or RNA. In some embodiments, the DNA or RNA measurements include measurements (such as the quantity of each allele at each locus) at a set of polymorphic loci on one or more chromosomes or chromosome segments of interest.

**[41]** In one aspect, the invention features methods for selecting a therapy for the treatment, stabilization, or prevention of a disease or disorder in a mammal. In some embodiments, the method includes determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment using any of the methods described herein. In some embodiments, a therapy is selected for the mammal (such as a therapy for a disease or disorder associated with the overrepresentation of the first homologous chromosome segment).

**[42]** In one aspect, the invention features methods for preventing, delaying, stabilizing, or treating a disease or disorder in a mammal. In some embodiments, the method includes determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment using any of the methods described herein. In some embodiments, a therapy is selected for the mammal (such as a therapy for a disease or disorder associated with the overrepresentation of the first homologous chromosome segment) and then the therapy is

administered to the mammal.

**[43]** In some embodiments, treating, stabilizing, or preventing a disease or disorder includes preventing or delaying an initial or subsequent occurrence of a disease or disorder, increasing the disease-free survival time between the disappearance of a condition and its reoccurrence, stabilizing or reducing an adverse symptom associated with a condition, or inhibiting or stabilizing the progression of a condition. In some embodiments, at least 20, 40, 60, 80, 90, or 95% of the treated subjects have a complete remission in which all evidence of the condition disappears. In some embodiments, the length of time a subject survives after being diagnosed with a condition and treated is at least 20, 40, 60, 80, 100, 200, or even 500% greater than (i) the average amount of time an untreated subject survives or (ii) the average amount of time a subject treated with another therapy survives.

**[44]** In some embodiments, treating, stabilizing, or preventing cancer includes reducing or stabilizing the size of a tumor (*e.g.*, a benign or malignant tumor), slowing or preventing an increase in the size of a tumor, reducing or stabilizing the number of tumor cells, increasing the disease-free survival time between the disappearance of a tumor and its reappearance, preventing an initial or subsequent occurrence of a tumor, or reducing or stabilizing an adverse symptom associated with a tumor. In one embodiment, the number of cancerous cells surviving the treatment is at least 10, 20, 40, 60, 80, or 100% lower than the initial number of cancerous cells, as measured using any standard assay. In some embodiments, the decrease in the number of cancerous cells induced by administration of a therapy of the invention is at least 2, 5, 10, 20, or 50-fold greater than the decrease in the number of non-cancerous cells. In some embodiments, the number of cancerous cells present after administration of a therapy is at least 2, 5, 10, 20, or 50-fold lower than the number of cancerous cells present after administration of a control (such as administration of saline or a buffer). In some embodiments, the methods of the present invention result in a decrease of 10, 20, 40, 60, 80, or 100% in the size of a tumor as determined using standard methods. In some embodiments, at least 10, 20, 40, 60, 80, 90, or 95% of the treated subjects have a complete remission in which there are no detectable cancerous cells. In some embodiments, the cancer does not reappear, or reappears after at least 2, 5, 10, 15, or 20 years. In some embodiments, the length of time a subject survives after being diagnosed with cancer and treated with a therapy of the invention is at least 10, 20, 40, 60, 80, 100, 200, or even 500% greater than (i) the average amount of time an untreated subject survives or (ii) the average amount of time a subject treated with another therapy survives.

**[45]** In one aspect, the invention features methods for stratification of subjects involved in a clinical trial for the treatment, stabilization, or prevention of a disease or disorder in a mammal. In some

administered to the mammal.

**[43]** In some embodiments, treating, stabilizing, or preventing a disease or disorder includes preventing or delaying an initial or subsequent occurrence of a disease or disorder, increasing the disease-free survival time between the disappearance of a condition and its reoccurrence, stabilizing or reducing an adverse symptom associated with a condition, or inhibiting or stabilizing the progression of a condition. In some embodiments, at least 20, 40, 60, 80, 90, or 95% of the treated subjects have a complete remission in which all evidence of the condition disappears. In some embodiments, the length of time a subject survives after being diagnosed with a condition and treated is at least 20, 40, 60, 80, 100, 200, or even 500% greater than (i) the average amount of time an untreated subject survives or (ii) the average amount of time a subject treated with another therapy survives.

**[44]** In some embodiments, treating, stabilizing, or preventing cancer includes reducing or stabilizing the size of a tumor (*e.g.*, a benign or malignant tumor), slowing or preventing an increase in the size of a tumor, reducing or stabilizing the number of tumor cells, increasing the disease-free survival time between the disappearance of a tumor and its reappearance, preventing an initial or subsequent occurrence of a tumor, or reducing or stabilizing an adverse symptom associated with a tumor. In one embodiment, the number of cancerous cells surviving the treatment is at least 10, 20, 40, 60, 80, or 100% lower than the initial number of cancerous cells, as measured using any standard assay. In some embodiments, the decrease in the number of cancerous cells induced by administration of a therapy of the invention is at least 2, 5, 10, 20, or 50-fold greater than the decrease in the number of non-cancerous cells. In some embodiments, the number of cancerous cells present after administration of a therapy is at least 2, 5, 10, 20, or 50-fold lower than the number of cancerous cells present after administration of a control (such as administration of saline or a buffer). In some embodiments, the methods of the present invention result in a decrease of 10, 20, 40, 60, 80, or 100% in the size of a tumor as determined using standard methods. In some embodiments, at least 10, 20, 40, 60, 80, 90, or 95% of the treated subjects have a complete remission in which there are no detectable cancerous cells. In some embodiments, the cancer does not reappear, or reappears after at least 2, 5, 10, 15, or 20 years. In some embodiments, the length of time a subject survives after being diagnosed with cancer and treated with a therapy of the invention is at least 10, 20, 40, 60, 80, 100, 200, or even 500% greater than (i) the average amount of time an untreated subject survives or (ii) the average amount of time a subject treated with another therapy survives.

**[45]** In one aspect, the invention features methods for stratification of subjects involved in a clinical trial for the treatment, stabilization, or prevention of a disease or disorder in a mammal. In some

embodiments, the method includes determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment using any of the methods described herein before, during, or after the clinical trial. In some embodiments, the presence or absence of the overrepresentation of the first homologous chromosome segment in the genome of the subject places the subject into a subgroup for the clinical trial.

[46] In some embodiments, the disease or disorder is selected from the group consisting of cancer, mental handicap, learning disability (*e.g.*, idiopathic learning disability), mental retardation, developmental delay, autism, neurodegenerative disease or disorder, schizophrenia, physical handicap, autoimmune disease or disorder, systemic lupus erythematosus, psoriasis, Crohn's disease, glomerulonephritis, HIV infection, AIDS, and combinations thereof. In some embodiments, the disease or disorder is selected from the group consisting of DiGeorge syndrome, DiGeorge 2 syndrome, DiGeorge/VCFS syndrome, Prader-Willi syndrome, Angelman syndrome, Beckwith-Wiedemann syndrome, 1p36 deletion syndrome, 2q37 deletion syndrome, 3q29 deletion syndrome, 9q34 deletion syndrome, 17q21.31 deletion syndrome, Cri-du-chat syndrome, Jacobsen syndrome, Miller Dieker syndrome, Phelan-McDermid syndrome, Smith-Magenis syndrome, WAGR syndrome, Wolf-Hirschhorn syndrome, Williams syndrome, Williams-Beuren syndrome, Miller-Dieker syndrome, Phelan-McDermid syndrome, Smith-Magenis syndrome, Down syndrome, Edward syndrome, Patau syndrome, Klinefelter syndrome, Turner syndrome, 47,XXX syndrome, 47,XYY syndrome, Sotos syndrome, and combinations thereof. In some embodiments, the method determines the presence or absence of one or more of the following chromosomal abnormalities: nullsomy, monosomy, uniparental disomy, trisomy, matched trisomy, unmatched trisomy, maternal trisomy, paternal trisomy, triploidy, mosaicism tetrasomy, matched tetrasomy, unmatched tetrasomy, other aneuploidies, unbalanced translocations, balanced translocations, insertions, deletions, recombinations, and combinations thereof. In some embodiments, the chromosomal abnormality is any deviation in the copy number of a specific chromosome or chromosome segment from the most common number of copies of that segment or chromosome, for example in a human somatic cell, any deviation from 2 copies can be regarded as a chromosomal abnormality. In some embodiments, the method determines the presence or absence of a euploidy. In some embodiments, the copy number hypotheses include one or more copy number hypotheses for a singleton pregnancy. In some embodiments, the copy number hypotheses include one or more copy number hypotheses for a multiple pregnancy, such as a twin pregnancy (*e.g.*, identical or fraternal twins or a vanishing twin). In some embodiments, the copy number hypotheses include all fetuses in a multiple pregnancy being euploid, all

embodiments, the method includes determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment using any of the methods described herein before, during, or after the clinical trial. In some embodiments, the presence or absence of the overrepresentation of the first homologous chromosome segment in the genome of the subject places the subject into a subgroup for the clinical trial.

[46] In some embodiments, the disease or disorder is selected from the group consisting of cancer, mental handicap, learning disability (*e.g.*, idiopathic learning disability), mental retardation, developmental delay, autism, neurodegenerative disease or disorder, schizophrenia, physical handicap, autoimmune disease or disorder, systemic lupus erythematosus, psoriasis, Crohn's disease, glomerulonephritis, HIV infection, AIDS, and combinations thereof. In some embodiments, the disease or disorder is selected from the group consisting of DiGeorge syndrome, DiGeorge 2 syndrome, DiGeorge/VCFS syndrome, Prader-Willi syndrome, Angelman syndrome, Beckwith-Wiedemann syndrome, 1p36 deletion syndrome, 2q37 deletion syndrome, 3q29 deletion syndrome, 9q34 deletion syndrome, 17q21.31 deletion syndrome, Cri-du-chat syndrome, Jacobsen syndrome, Miller Dieker syndrome, Phelan-McDermid syndrome, Smith-Magenis syndrome, WAGR syndrome, Wolf-Hirschhorn syndrome, Williams syndrome, Williams-Beuren syndrome, Miller-Dieker syndrome, Phelan-McDermid syndrome, Smith-Magenis syndrome, Down syndrome, Edward syndrome, Patau syndrome, Klinefelter syndrome, Turner syndrome, 47,XXX syndrome, 47,XYY syndrome, Sotos syndrome, and combinations thereof. In some embodiments, the method determines the presence or absence of one or more of the following chromosomal abnormalities: nullsomy, monosomy, uniparental disomy, trisomy, matched trisomy, unmatched trisomy, maternal trisomy, paternal trisomy, triploidy, mosaicism tetrasomy, matched tetrasomy, unmatched tetrasomy, other aneuploidies, unbalanced translocations, balanced translocations, insertions, deletions, recombinations, and combinations thereof. In some embodiments, the chromosomal abnormality is any deviation in the copy number of a specific chromosome or chromosome segment from the most common number of copies of that segment or chromosome, for example in a human somatic cell, any deviation from 2 copies can be regarded as a chromosomal abnormality. In some embodiments, the method determines the presence or absence of a euploidy. In some embodiments, the copy number hypotheses include one or more copy number hypotheses for a singleton pregnancy. In some embodiments, the copy number hypotheses include one or more copy number hypotheses for a multiple pregnancy, such as a twin pregnancy (*e.g.*, identical or fraternal twins or a vanishing twin). In some embodiments, the copy number hypotheses include all fetuses in a multiple pregnancy being euploid, all

fetuses in a multiple pregnancy being aneuploid (such as any of the aneuploidies disclosed herein), and/or one or more fetuses in a multiple pregnancy being euploid and one or more fetuses in a multiple pregnancy being aneuploidy. In some embodiments, the copy number hypotheses include identical twins (also referred to as monozygotic twins) or fraternal twins (also referred to as dizygotic twins). In some embodiments, the copy number hypotheses include a molar pregnancy, such as a complete or partial molar pregnancy. In some embodiments, the chromosome segment of interest is an entire chromosome. In some embodiments, the chromosome or chromosome segment is selected from the group consisting of chromosome 13, chromosome 18, chromosome 21, the X chromosome, the Y chromosome, segments thereof, and combinations thereof. In some embodiments, the first homologous chromosome segment and second homologous chromosome segment are a pair of homologous chromosome segments that comprises the chromosome segment of interest. In some embodiments, the first homologous chromosome segment and second homologous chromosome segment are a pair of homologous chromosomes of interest. In some embodiments, a confidence is computed for the CNV determination or the diagnosis of the disease or disorder.

**[47]** In some embodiments, the deletion is a deletion of at least 0.01 kb, 0.1 kb, 1 kb, 10 kb, 100 kb, 1 mb, 2 mb, 3 mb, 5 mb, 10 mb, 15 mb, 20 mb, 30 mb, or 40 mb. In some embodiments, the deletion is a deletion of between 1 kb to 40 mb, such as between 1 kb to 100 kb, 100 kb to 1 mb, 1 to 5 mb, 5 to 10 mb, 10 to 15 mb, 15 to 20 mb, 20 to 25 mb, 25 to 30 mb, or 30 to 40 mb, inclusive. In some embodiments, one copy of the chromosome segment is deleted and one copy is present. In some embodiments, two copies of the chromosome segment are deleted. In some embodiments, an entire chromosome is deleted.

**[48]** In some embodiments, the duplication is a duplication of at least 0.01 kb, 0.1 kb, 1 kb, 10 kb, 100 kb, 1 mb, 2 mb, 3 mb, 5 mb, 10 mb, 15 mb, 20 mb, 30 mb, or 40 mb. In some embodiments, the duplication is a duplication of between 1 kb to 40 mb, such as between 1 kb to 100 kb, 100 kb to 1 mb, 1 to 5 mb, 5 to 10 mb, 10 to 15 mb, 15 to 20 mb, 20 to 25 mb, 25 to 30 mb, or 30 to 40 mb, inclusive. In some embodiments, the chromosome segment is duplicated one time. In some embodiments, the chromosome segment is duplicated more than one time, such as 2, 3, 4, or 5 times. In some embodiments, an entire chromosome is duplicated. In some embodiments, a region in a first homologous segment is deleted, and the same region or another region in the second homologous segment is duplicated. In some embodiments, at least 50, 60, 70, 80, 90, 95, 96, 98, 99, or 100% of the SNVs tested for are transversion mutations rather than transition mutations.

**[49]** In some embodiments, the sample comprises DNA and/or RNA from (i) one or more target cells

fetuses in a multiple pregnancy being aneuploid (such as any of the aneuploidies disclosed herein), and/or one or more fetuses in a multiple pregnancy being euploid and one or more fetuses in a multiple pregnancy being aneuploidy. In some embodiments, the copy number hypotheses include identical twins (also referred to as monozygotic twins) or fraternal twins (also referred to as dizygotic twins). In some embodiments, the copy number hypotheses include a molar pregnancy, such as a complete or partial molar pregnancy. In some embodiments, the chromosome segment of interest is an entire chromosome. In some embodiments, the chromosome or chromosome segment is selected from the group consisting of chromosome 13, chromosome 18, chromosome 21, the X chromosome, the Y chromosome, segments thereof, and combinations thereof. In some embodiments, the first homologous chromosome segment and second homologous chromosome segment are a pair of homologous chromosome segments that comprises the chromosome segment of interest. In some embodiments, the first homologous chromosome segment and second homologous chromosome segment are a pair of homologous chromosomes of interest. In some embodiments, a confidence is computed for the CNV determination or the diagnosis of the disease or disorder.

**[47]** In some embodiments, the deletion is a deletion of at least 0.01 kb, 0.1 kb, 1 kb, 10 kb, 100 kb, 1 mb, 2 mb, 3 mb, 5 mb, 10 mb, 15 mb, 20 mb, 30 mb, or 40 mb. In some embodiments, the deletion is a deletion of between 1 kb to 40 mb, such as between 1 kb to 100 kb, 100 kb to 1 mb, 1 to 5 mb, 5 to 10 mb, 10 to 15 mb, 15 to 20 mb, 20 to 25 mb, 25 to 30 mb, or 30 to 40 mb, inclusive. In some embodiments, one copy of the chromosome segment is deleted and one copy is present. In some embodiments, two copies of the chromosome segment are deleted. In some embodiments, an entire chromosome is deleted.

**[48]** In some embodiments, the duplication is a duplication of at least 0.01 kb, 0.1 kb, 1 kb, 10 kb, 100 kb, 1 mb, 2 mb, 3 mb, 5 mb, 10 mb, 15 mb, 20 mb, 30 mb, or 40 mb. In some embodiments, the duplication is a duplication of between 1 kb to 40 mb, such as between 1 kb to 100 kb, 100 kb to 1 mb, 1 to 5 mb, 5 to 10 mb, 10 to 15 mb, 15 to 20 mb, 20 to 25 mb, 25 to 30 mb, or 30 to 40 mb, inclusive. In some embodiments, the chromosome segment is duplicated one time. In some embodiments, the chromosome segment is duplicated more than one time, such as 2, 3, 4, or 5 times. In some embodiments, an entire chromosome is duplicated. In some embodiments, a region in a first homologous segment is deleted, and the same region or another region in the second homologous segment is duplicated. In some embodiments, at least 50, 60, 70, 80, 90, 95, 96, 98, 99, or 100% of the SNVs tested for are transversion mutations rather than transition mutations.

**[49]** In some embodiments, the sample comprises DNA and/or RNA from (i) one or more target cells

or (ii) one or more non-target cells. In some embodiments, the sample is a mixed sample with DNA and/or RNA from one or more target cells and one or more non-target cells. In some embodiments, the target cells are cells that have a CNV, such as a deletion or duplication of interest, and the non-target cells are cells that do not have the copy number variation of interest. In some embodiments in which the one or more target cells are cancer cell(s) and the one or more non-target cells are non-cancerous cell(s), the method includes determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more of the cancer cells. In some embodiments in which the one or more target cells are genetically identical cancer cell(s) and the one or more non-target cells are non-cancerous cell(s), the method includes determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of the cancer cell(s). In some embodiments in which the one or more target cells are genetically non-identical cancer cell(s) and the one or more non-target cells are non-cancerous cell(s), the method includes determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more of the genetically non-identical cancer cells. In some embodiments in which the sample comprises cell-free DNA from a mixture of one or more cancer cells and one or more non-cancerous cells, the method includes determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more of the cancer cells. In some embodiments in which the one or more target cells are genetically identical fetal cell(s) and the one or more non-target cells are maternal cell(s), the method includes determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of the fetal cell(s). In some embodiments in which the one or more target cells are genetically non-identical fetal cell(s) and the one or more non-target cells are maternal cell(s), the method includes determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more of the genetically nonidentical fetal cells. As the cells of most individuals contain a nearly identical set of nuclear DNA, the term “target cell” may be used interchangeably with the term “individual” in some embodiments. Cancerous cells have genotypes that are distinct from the host individual. In this case, the cancer itself may be considered an individual. Moreover, many cancers are heterogeneous meaning that different cells in a tumor are genetically distinct from other cells in the same tumor. In this case, the different genetically identical regions can be considered different individuals. Alternately, the cancer may be considered a single individual with a mixture of cells with distinct genomes. Typically, non-target cells are euploid, though this is not necessarily the case.

or (ii) one or more non-target cells. In some embodiments, the sample is a mixed sample with DNA and/or RNA from one or more target cells and one or more non-target cells. In some embodiments, the target cells are cells that have a CNV, such as a deletion or duplication of interest, and the non-target cells are cells that do not have the copy number variation of interest. In some embodiments in which the one or more target cells are cancer cell(s) and the one or more non-target cells are non-cancerous cell(s), the method includes determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more of the cancer cells. In some embodiments in which the one or more target cells are genetically identical cancer cell(s) and the one or more non-target cells are non-cancerous cell(s), the method includes determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of the cancer cell(s). In some embodiments in which the one or more target cells are genetically non-identical cancer cell(s) and the one or more non-target cells are non-cancerous cell(s), the method includes determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more of the genetically non-identical cancer cells. In some embodiments in which the sample comprises cell-free DNA from a mixture of one or more cancer cells and one or more non-cancerous cells, the method includes determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more of the cancer cells. In some embodiments in which the one or more target cells are genetically identical fetal cell(s) and the one or more non-target cells are maternal cell(s), the method includes determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of the fetal cell(s). In some embodiments in which the one or more target cells are genetically non-identical fetal cell(s) and the one or more non-target cells are maternal cell(s), the method includes determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more of the genetically nonidentical fetal cells. As the cells of most individuals contain a nearly identical set of nuclear DNA, the term “target cell” may be used interchangeably with the term “individual” in some embodiments. Cancerous cells have genotypes that are distinct from the host individual. In this case, the cancer itself may be considered an individual. Moreover, many cancers are heterogeneous meaning that different cells in a tumor are genetically distinct from other cells in the same tumor. In this case, the different genetically identical regions can be considered different individuals. Alternately, the cancer may be considered a single individual with a mixture of cells with distinct genomes. Typically, non-target cells are euploid, though this is not necessarily the case.

**[50]** In some embodiments, the sample is obtained from a maternal whole blood sample or fraction thereof, cells isolated from a maternal blood sample, an amniocentesis sample, a products of conception sample, a placental tissue sample, a chorionic villus sample, a placental membrane sample, a cervical mucus sample, or a sample from a fetus. In some embodiments, the sample comprises cell-free DNA obtained from a blood sample or fraction thereof from the mother. In some embodiments, the sample comprises nuclear DNA obtained from a mixture of fetal cells and maternal cells. In some embodiments, the sample is obtained from a fraction of maternal blood containing nucleated cells that has been enriched for fetal cells. In some embodiments, a sample is divided into multiple fractions (such as 2, 3, 4 5, or more fractions) that are each analyzed using a method of the invention. If each fraction produces the same results (such as the presence or absence of one or more CNVs of interest), the confidence in the results increases. In different fractions produce different results, the sample could be re-analyzed or another sample could be collected from the same subject and analyzed.

**[51]** Exemplary subjects include mammals, such as humans and mammals of veterinary interest. In some embodiments, the mammal is a primate (*e.g.*, a human, a monkey, a gorilla, an ape, a lemur, *etc.*), a bovine, an equine, a porcine, a canine, or a feline.

**[52]** In some embodiments, any of the methods include generating a report (such as a written or electronic report) disclosing a result of the method of the invention (such as the presence or absence of a deletion or duplication).

**[53]** In some embodiments, any of the methods include taking a clinical action based on a result of a method of the invention (such as the presence or absence of a deletion or duplication). In some embodiments in which an embryo or fetus has one or more polymorphisms or mutations of interest (such as a CNV) based on a result of a method of the invention, the clinical action includes performing additional testing (such as testing to confirm the presence of the polymorphism or mutation), not implanting the embryo for IVF, implanting a different embryo for IVF, terminating a pregnancy, preparing for a special needs child, or undergoing an intervention designed to decrease the severity of the phenotypic presentation of a genetic disorder. In some embodiments, the clinical action is selected from the group consisting of performing an ultrasound, amniocentesis on the fetus, amniocentesis on a subsequent fetus that inherits genetic material from the mother and/or father, chorion villus biopsy on the fetus, chorion villus biopsy on a subsequent fetus that inherits genetic material from the mother and/or father, *in vitro* fertilization, preimplantation genetic diagnosis on one or more embryos that inherited genetic material from the mother and/or father, karyotyping on the mother, karyotyping on the father, fetal echocardiogram (such as an

**[50]** In some embodiments, the sample is obtained from a maternal whole blood sample or fraction thereof, cells isolated from a maternal blood sample, an amniocentesis sample, a products of conception sample, a placental tissue sample, a chorionic villus sample, a placental membrane sample, a cervical mucus sample, or a sample from a fetus. In some embodiments, the sample comprises cell-free DNA obtained from a blood sample or fraction thereof from the mother. In some embodiments, the sample comprises nuclear DNA obtained from a mixture of fetal cells and maternal cells. In some embodiments, the sample is obtained from a fraction of maternal blood containing nucleated cells that has been enriched for fetal cells. In some embodiments, a sample is divided into multiple fractions (such as 2, 3, 4, 5, or more fractions) that are each analyzed using a method of the invention. If each fraction produces the same results (such as the presence or absence of one or more CNVs of interest), the confidence in the results increases. In different fractions produce different results, the sample could be re-analyzed or another sample could be collected from the same subject and analyzed.

**[51]** Exemplary subjects include mammals, such as humans and mammals of veterinary interest. In some embodiments, the mammal is a primate (*e.g.*, a human, a monkey, a gorilla, an ape, a lemur, *etc.*), a bovine, an equine, a porcine, a canine, or a feline.

**[52]** In some embodiments, any of the methods include generating a report (such as a written or electronic report) disclosing a result of the method of the invention (such as the presence or absence of a deletion or duplication).

**[53]** In some embodiments, any of the methods include taking a clinical action based on a result of a method of the invention (such as the presence or absence of a deletion or duplication). In some embodiments in which an embryo or fetus has one or more polymorphisms or mutations of interest (such as a CNV) based on a result of a method of the invention, the clinical action includes performing additional testing (such as testing to confirm the presence of the polymorphism or mutation), not implanting the embryo for IVF, implanting a different embryo for IVF, terminating a pregnancy, preparing for a special needs child, or undergoing an intervention designed to decrease the severity of the phenotypic presentation of a genetic disorder. In some embodiments, the clinical action is selected from the group consisting of performing an ultrasound, amniocentesis on the fetus, amniocentesis on a subsequent fetus that inherits genetic material from the mother and/or father, chorion villus biopsy on the fetus, chorion villus biopsy on a subsequent fetus that inherits genetic material from the mother and/or father, *in vitro* fertilization, preimplantation genetic diagnosis on one or more embryos that inherited genetic material from the mother and/or father, karyotyping on the mother, karyotyping on the father, fetal echocardiogram (such as an

echocardiogram of a fetus with trisomy 21, 18, or 13, monosomy X, or a microdeletion) and combinations thereof. In some embodiments, the clinical action is selected from the group consisting of administering growth hormone to a bom child with monosomy X (such as administration starting at ~9 months), administering calcium to a bom child with a 22q deletion (such as DiGeorge syndrome), administering an androgen such as testosterone to a bom child with 47,XXY (such as one injection per month for 3 months of 25 mg testosterone enanthate to an infant or toddler), performing a test for cancer on a woman with a complete or partial molar pregnancy (such as a triploid fetus), administering a therapy for cancer such as a chemotherapeutic agent to a woman with a complete or partial molar pregnancy (such as a triploid fetus), screening a fetus determined to be male (such as a fetus determined to be male using a method of the invention) for one or more X-linked genetic disorders such as

Duchenne muscular dystrophy (DMD), adrenoleukodystrophy, or hemophilia, performing amniocentesis on a male fetus at risk for an X-linked disorder, administering dexamethasone to a women with a female fetus (such as a fetus determined to be female using a method of the invention) at risk for congenital adrenal hyperplasia, performing amniocentesis on a female fetus at risk for congenital adrenal hyperplasia, administering killed vaccines (instead of live vaccines) or not administering certain vaccines to a bom child that is (or is suspected of being) immune deficient from a 22ql.2 deletion, performing occupational and/or physical therapy, performing early intervention in education, delivering the baby at a tertiary care center with a NICU and/or having pediatric specialists available at delivery, behavioral intervention for bom child (such as a child with XXX, XXY, or XYY), and combinations thereof.

**[54]** In some embodiments, ultrasound or another screening test is performed on a women determined to have multiple pregnancies (such as twins) to determine whether or not two or more of the fetus are monochorionic. Monozygotic twins result from ovulation and fertilization of a single oocyte, with subsequent division of the zygote; placentation may be dichorionic or monochorionic. Dizygotic twins occur from ovulation and fertilization of two oocytes, which usually results in dichorionic placentation. Monochorionic twins have a risk of twin-to-twin transfusion syndrome, which may cause unequal distribution of blood between fetuses that results in differences in their growth and development, sometimes resulting in stillbirth. Thus, twins determined to be monozygotic twins using a method of the invention are desirably tested (such as by ultrasound) to determine if they are monochorionic twins, and if so, these twins can be monitored (such as bi-weekly ultrasounds from 16 weeks) for signs of win-to-twin transfusion syndrome.

**[55]** In some embodiments in which an embryo or fetus does not have one or more one or more

echocardiogram of a fetus with trisomy 21, 18, or 13, monosomy X, or a microdeletion) and combinations thereof. In some embodiments, the clinical action is selected from the group consisting of administering growth hormone to a bom child with monosomy X (such as administration starting at ~9 months), administering calcium to a bom child with a 22q deletion (such as DiGeorge syndrome), administering an androgen such as testosterone to a bom child with 47,XXY (such as one injection per month for 3 months of 25 mg testosterone enanthate to an infant or toddler), performing a test for cancer on a woman with a complete or partial molar pregnancy (such as a triploid fetus), administering a therapy for cancer such as a chemotherapeutic agent to a woman with a complete or partial molar pregnancy (such as a triploid fetus), screening a fetus determined to be male (such as a fetus determined to be male using a method of the invention) for one or more X-linked genetic disorders such as

Duchenne muscular dystrophy (DMD), adrenoleukodystrophy, or hemophilia, performing amniocentesis on a male fetus at risk for an X-linked disorder, administering dexamethasone to a women with a female fetus (such as a fetus determined to be female using a method of the invention) at risk for congenital adrenal hyperplasia, performing amniocentesis on a female fetus at risk for congenital adrenal hyperplasia, administering killed vaccines (instead of live vaccines) or not administering certain vaccines to a bom child that is (or is suspected of being) immune deficient from a 22ql.2 deletion, performing occupational and/or physical therapy, performing early intervention in education, delivering the baby at a tertiary care center with a NICU and/or having pediatric specialists available at delivery, behavioral intervention for bom child (such as a child with XXX, XXY, or XYY), and combinations thereof.

**[54]** In some embodiments, ultrasound or another screening test is performed on a women determined to have multiple pregnancies (such as twins) to determine whether or not two or more of the fetus are monochorionic. Monozygotic twins result from ovulation and fertilization of a single oocyte, with subsequent division of the zygote; placentation may be dichorionic or monochorionic. Dizygotic twins occur from ovulation and fertilization of two oocytes, which usually results in dichorionic placentation. Monochorionic twins have a risk of twin-to-twin transfusion syndrome, which may cause unequal distribution of blood between fetuses that results in differences in their growth and development, sometimes resulting in stillbirth. Thus, twins determined to be monozygotic twins using a method of the invention are desirably tested (such as by ultrasound) to determine if they are monochorionic twins, and if so, these twins can be monitored (such as bi-weekly ultrasounds from 16 weeks) for signs of win-to-twin transfusion syndrome.

**[55]** In some embodiments in which an embryo or fetus does not have one or more one or more

polymorphisms or mutations of interest (such as a CNV) based on a result of a method of the invention, the clinical action includes implanting the embryo for IVF or continuing a pregnancy. In some embodiments, the clinical action is additional testing to confirm the absence of the polymorphism or mutation selected from the group consisting of performing an ultrasound, amniocentesis, chorion villus biopsy, and combinations thereof.

**[56]** In some embodiments in which an individual has one or more one or more polymorphisms or mutations (such as a polymorphism or mutation associated with a disease or disorder such as cancer or an increased risk for a disease or disorder such as cancer) based on a result of a method of the invention, the clinical action includes performing additional testing or administering one or more therapies for a disease or disorder (such as a therapy for cancer, a therapy for the specific type of cancer or type of mutation the individual is diagnosed with, or any of the therapies disclosed herein). In some embodiments, the clinical action is additional testing to confirm the presence or absence of a polymorphism or mutation selected from the group consisting of biopsy, surgery, medical imaging (such as a mammogram or an ultrasound), and combinations thereof.

**[57]** In some embodiments, the additional testing includes performing the same or a different method (such as any of the methods described herein) to confirm the presence or absence of the polymorphism or mutation (such as a CNV), such as testing either a second fraction of the same sample that was tested or a different sample from the same individual (such as the same pregnant mother, fetus, embryo, or individual at increased risk for cancer). In some embodiments, the additional testing is performed for an individual for whom the probability of a polymorphism or mutation (such as a CNV) is above a threshold value (such as additional testing to confirm the presence of a likely polymorphism or mutation). In some embodiments, the additional testing is performed for an individual for whom the confidence or z-score for the determination of a polymorphism or mutation (such as a CNV) is above a threshold value (such as additional testing to confirm the presence of a likely polymorphism or mutation). In some embodiments, the additional testing is performed for an individual for whom the confidence or z-score for the determination of a polymorphism or mutation (such as a CNV) is between minimum and maximum threshold values (such as additional testing to increase the confidence that the initial result is correct). In some embodiments, the additional testing is performed for an individual for whom the confidence for the determination of the presence or absence of a polymorphism or mutation (such as a CNV) is below a threshold value (such as a “no call” result due to not being able to determine the presence or absence of the CNV with sufficient confidence). An exemplary Z core is calculated in Chiu et al. BMJ 201

polymorphisms or mutations of interest (such as a CNV) based on a result of a method of the invention, the clinical action includes implanting the embryo for IVF or continuing a pregnancy. In some embodiments, the clinical action is additional testing to confirm the absence of the polymorphism or mutation selected from the group consisting of performing an ultrasound, amniocentesis, chorion villus biopsy, and combinations thereof.

**[56]** In some embodiments in which an individual has one or more one or more polymorphisms or mutations (such as a polymorphism or mutation associated with a disease or disorder such as cancer or an increased risk for a disease or disorder such as cancer) based on a result of a method of the invention, the clinical action includes performing additional testing or administering one or more therapies for a disease or disorder (such as a therapy for cancer, a therapy for the specific type of cancer or type of mutation the individual is diagnosed with, or any of the therapies disclosed herein). In some embodiments, the clinical action is additional testing to confirm the presence or absence of a polymorphism or mutation selected from the group consisting of biopsy, surgery, medical imaging (such as a mammogram or an ultrasound), and combinations thereof.

**[57]** In some embodiments, the additional testing includes performing the same or a different method (such as any of the methods described herein) to confirm the presence or absence of the polymorphism or mutation (such as a CNV), such as testing either a second fraction of the same sample that was tested or a different sample from the same individual (such as the same pregnant mother, fetus, embryo, or individual at increased risk for cancer). In some embodiments, the additional testing is performed for an individual for whom the probability of a polymorphism or mutation (such as a CNV) is above a threshold value (such as additional testing to confirm the presence of a likely polymorphism or mutation). In some embodiments, the additional testing is performed for an individual for whom the confidence or z-score for the determination of a polymorphism or mutation (such as a CNV) is above a threshold value (such as additional testing to confirm the presence of a likely polymorphism or mutation). In some embodiments, the additional testing is performed for an individual for whom the confidence or z-score for the determination of a polymorphism or mutation (such as a CNV) is between minimum and maximum threshold values (such as additional testing to increase the confidence that the initial result is correct). In some embodiments, the additional testing is performed for an individual for whom the confidence for the determination of the presence or absence of a polymorphism or mutation (such as a CNV) is below a threshold value (such as a “no call” result due to not being able to determine the presence or absence of the CNV with sufficient confidence). An exemplary Z core is calculated in Chiu et al. BMJ 201

1;342:c7401 (which is hereby incorporated by reference in its entirety) in which chromosome 21 is used as an example and can be replaced with any other chromosome or chromosome segment in the test sample.

Z score for percentage chromosome 21 in test case = ((percentage chromosome 21 in test case) - (mean percentage chromosome 21 in reference controls)) / (standard deviation of percentage chromosome 21 in reference controls).

In some embodiments, the additional testing is performed for an individual for whom the initial sample did not meet quality control guidelines or had a fetal fraction or a tumor fraction below a threshold value. In some embodiments, the method includes selecting an individual for additional testing based on the result of a method of the invention, the probability of the result, the confidence of the result, or the z-score; and performing the additional testing on the individual (such as on the same or a different sample). In some embodiments, a subject diagnosed with a disease or disorder (such as cancer) undergoes repeat testing using a method of the invention or known testing for the disease or disorder at multiple time points to monitor the progression of the disease or disorder or the remission or reoccurrence of the disease or disorder.

**[58]** In one aspect, the invention features a report (such as a written or electronic report) with a result from a method of the invention (such as the presence or absence of a deletion or duplication).

**[59]** In various embodiments, the primer extension reaction or the polymerase chain reaction includes the addition of one or more nucleotides by a polymerase. In some embodiments, the primers are in solution. In some embodiments, the primers are in solution and are not immobilized on a solid support. In some embodiments, the primers are not part of a microarray. In various embodiments, the primer extension reaction or the polymerase chain reaction does not include ligation-mediated PCR. In various embodiments, the primer extension reaction or the polymerase chain reaction does not include the joining of two primers by a ligase. In various embodiments, the primers do not include Linked Inverted Probes (LIPs), which can also be called pre-circularized probes, pre-circularizing probes, circularizing probes, Padlock Probes, or Molecular Inversion Probes (MIPs).

**[60]** It is understood that aspects and embodiments of the invention described herein include “comprising,” “consisting,” and “consisting essentially of” aspects and embodiments. It is understood that aspects and embodiments of the invention described herein include combinations of any two or more of the aspects or embodiments of the invention.

### *Definitions*

1;342:c7401 (which is hereby incorporated by reference in its entirety) in which chromosome 21 is used as an example and can be replaced with any other chromosome or chromosome segment in the test sample.

Z score for percentage chromosome 21 in test case = ((percentage chromosome 21 in test case) - (mean percentage chromosome 21 in reference controls)) / (standard deviation of percentage chromosome 21 in reference controls).

In some embodiments, the additional testing is performed for an individual for whom the initial sample did not meet quality control guidelines or had a fetal fraction or a tumor fraction below a threshold value. In some embodiments, the method includes selecting an individual for additional testing based on the result of a method of the invention, the probability of the result, the confidence of the result, or the z-score; and performing the additional testing on the individual (such as on the same or a different sample). In some embodiments, a subject diagnosed with a disease or disorder (such as cancer) undergoes repeat testing using a method of the invention or known testing for the disease or disorder at multiple time points to monitor the progression of the disease or disorder or the remission or reoccurrence of the disease or disorder.

**[58]** In one aspect, the invention features a report (such as a written or electronic report) with a result from a method of the invention (such as the presence or absence of a deletion or duplication).

**[59]** In various embodiments, the primer extension reaction or the polymerase chain reaction includes the addition of one or more nucleotides by a polymerase. In some embodiments, the primers are in solution. In some embodiments, the primers are in solution and are not immobilized on a solid support. In some embodiments, the primers are not part of a microarray. In various embodiments, the primer extension reaction or the polymerase chain reaction does not include ligation-mediated PCR. In various embodiments, the primer extension reaction or the polymerase chain reaction does not include the joining of two primers by a ligase. In various embodiments, the primers do not include Linked Inverted Probes (LIPs), which can also be called pre-circularized probes, pre-circularizing probes, circularizing probes, Padlock Probes, or Molecular Inversion Probes (MIPs).

**[60]** It is understood that aspects and embodiments of the invention described herein include “comprising,” “consisting,” and “consisting essentially of” aspects and embodiments. It is understood that aspects and embodiments of the invention described herein include combinations of any two or more of the aspects or embodiments of the invention.

### *Definitions*

[61] *Single Nucleotide Polymorphism (SNP)* refers to a single nucleotide that may differ between the genomes of two members of the same species. The usage of the term should not imply any limit on the frequency with which each variant occurs.

[62] *Sequence* refers to a DNA sequence or a genetic sequence. It may refer to the primary, physical structure of the DNA molecule or strand in an individual. It may refer to the sequence of nucleotides found in that DNA molecule, or the complementary strand to the DNA molecule. It may refer to the information contained in the DNA molecule as its representation *in silico*.

[63] *Locus* refers to a particular region of interest on the DNA of an individual, which may refer to a SNP, the site of a possible insertion or deletion, or the site of some other relevant genetic variation. Disease-linked SNPs may also refer to disease-linked loci.

[64] *Polymorphic Allele*, also “*Polymorphic Locus*,” refers to an allele or locus where the genotype varies between individuals within a given species. Some examples of polymorphic alleles include single nucleotide polymorphisms, short tandem repeats, deletions, duplications, and inversions.

[65] *Polymorphic Site* refers to the specific nucleotides found in a polymorphic region that vary between individuals.

[66] *Mutation* refers to an alteration in a naturally-occurring or reference nucleic acid sequence, such as an insertion, deletion, duplication, translocation, substitution, frameshift mutation, silent mutation, nonsense mutation, missense mutation, point mutation, transition mutation, transversion mutation, reverse mutation, or microsatellite alteration. In some embodiments, the amino acid sequence encoded by the nucleic acid sequence has at least one amino acid alteration from a naturally-occurring sequence.

[67] *Allele* refers to the genes that occupy a particular locus.

[68] *Genetic Data* also “*Genotypic Data*” refers to the data describing aspects of the genome of one or more individuals. It may refer to one or a set of loci, partial or entire sequences, partial or entire chromosomes, or the entire genome. It may refer to the identity of one or a plurality of nucleotides; it may refer to a set of sequential nucleotides, or nucleotides from different locations in the genome, or a combination thereof. Genotypic data is typically *in silico*, however, it is also possible to consider physical nucleotides in a sequence as chemically encoded genetic data. Genotypic Data may be said to be “on,” “of,” “at,” “from” or “on” the individual(s). Genotypic

Data may refer to output measurements from a genotyping platform where those measurements are made on genetic material.

[69] *Genetic Material* also “*Genetic Sample*” refers to physical matter, such as tissue or blood, from

[61] *Single Nucleotide Polymorphism (SNP)* refers to a single nucleotide that may differ between the genomes of two members of the same species. The usage of the term should not imply any limit on the frequency with which each variant occurs.

[62] *Sequence* refers to a DNA sequence or a genetic sequence. It may refer to the primary, physical structure of the DNA molecule or strand in an individual. It may refer to the sequence of nucleotides found in that DNA molecule, or the complementary strand to the DNA molecule. It may refer to the information contained in the DNA molecule as its representation *in silico*.

[63] *Locus* refers to a particular region of interest on the DNA of an individual, which may refer to a SNP, the site of a possible insertion or deletion, or the site of some other relevant genetic variation. Disease-linked SNPs may also refer to disease-linked loci.

[64] *Polymorphic Allele*, also “*Polymorphic Locus*,” refers to an allele or locus where the genotype varies between individuals within a given species. Some examples of polymorphic alleles include single nucleotide polymorphisms, short tandem repeats, deletions, duplications, and inversions.

[65] *Polymorphic Site* refers to the specific nucleotides found in a polymorphic region that vary between individuals.

[66] *Mutation* refers to an alteration in a naturally-occurring or reference nucleic acid sequence, such as an insertion, deletion, duplication, translocation, substitution, frameshift mutation, silent mutation, nonsense mutation, missense mutation, point mutation, transition mutation, transversion mutation, reverse mutation, or microsatellite alteration. In some embodiments, the amino acid sequence encoded by the nucleic acid sequence has at least one amino acid alteration from a naturally-occurring sequence.

[67] *Allele* refers to the genes that occupy a particular locus.

[68] *Genetic Data* also “*Genotypic Data*” refers to the data describing aspects of the genome of one or more individuals. It may refer to one or a set of loci, partial or entire sequences, partial or entire chromosomes, or the entire genome. It may refer to the identity of one or a plurality of nucleotides; it may refer to a set of sequential nucleotides, or nucleotides from different locations in the genome, or a combination thereof. Genotypic data is typically *in silico*, however, it is also possible to consider physical nucleotides in a sequence as chemically encoded genetic data. Genotypic Data may be said to be “on,” “of,” “at,” “from” or “on” the individual(s). Genotypic

Data may refer to output measurements from a genotyping platform where those measurements are made on genetic material.

[69] *Genetic Material* also “*Genetic Sample*” refers to physical matter, such as tissue or blood, from

one or more individuals comprising DNA or RNA.

**[70]** *Confidence* refers to the statistical likelihood that the called SNP, allele, set of alleles, determined number of copies of a chromosome or chromosome segment, or diagnosis of the presence or absence of a disease correctly represents the real genetic state of the individual.

**[71]** *Ploidy Calling*, also “Chromosome Copy Number Calling,” or “Copy Number Calling” (CNC), may refer to the act of determining the quantity and/or chromosomal identity of one or more chromosomes or chromosome segments present in a cell.

**[72]** *Aneuploidy* refers to the state where the wrong number of chromosomes (*e.g.*, the wrong number of full chromosomes or the wrong number of chromosome segments, such as the presence of deletions or duplications of a chromosome segment) is present in a cell. In the case of a somatic human cell it may refer to the case where a cell does not contain 22 pairs of autosomal chromosomes and one pair of sex chromosomes. In the case of a human gamete, it may refer to the case where a cell does not contain one of each of the 23 chromosomes. In the case of a single chromosome type, it may refer to the case where more or less than two homologous but non-identical chromosome copies are present, or where there are two chromosome copies present that originate from the same parent. In some embodiments, the deletion of a chromosome segment is a microdeletion.

**[73]** *Ploidy State* refers to the quantity and/or chromosomal identity of one or more chromosomes or chromosome segments in a cell.

**[74]** *Chromosome* may refer to a single chromosome copy, meaning a single molecule of DNA of which there are 46 in a normal somatic cell; an example is ‘the maternally derived chromosome 18’. Chromosome may also refer to a chromosome type, of which there are 23 in a normal human somatic cell; an example is ‘chromosome 18’.

**[75]** *Chromosomal Identity* may refer to the referent chromosome number, *i.e.* the chromosome type. Normal humans have 22 types of numbered autosomal chromosome types, and two types of sex chromosomes. It may also refer to the parental origin of the chromosome. It may also refer to a specific chromosome inherited from the parent. It may also refer to other identifying features of a chromosome.

**[76]** *Allelic Data* refers to a set of genotypic data concerning a set of one or more alleles. It may refer to the phased, haplotypic data. It may refer to SNP identities, and it may refer to the sequence data of the DNA, including insertions, deletions, repeats and mutations. It may include the parental origin of each allele.

**[77]** *Allelic State* refers to the actual state of the genes in a set of one or more alleles. It may refer to

one or more individuals comprising DNA or RNA.

**[70]** *Confidence* refers to the statistical likelihood that the called SNP, allele, set of alleles, determined number of copies of a chromosome or chromosome segment, or diagnosis of the presence or absence of a disease correctly represents the real genetic state of the individual.

**[71]** *Ploidy Calling*, also “Chromosome Copy Number Calling,” or “Copy Number Calling” (CNC), may refer to the act of determining the quantity and/or chromosomal identity of one or more chromosomes or chromosome segments present in a cell.

**[72]** *Aneuploidy* refers to the state where the wrong number of chromosomes (*e.g.*, the wrong number of full chromosomes or the wrong number of chromosome segments, such as the presence of deletions or duplications of a chromosome segment) is present in a cell. In the case of a somatic human cell it may refer to the case where a cell does not contain 22 pairs of autosomal chromosomes and one pair of sex chromosomes. In the case of a human gamete, it may refer to the case where a cell does not contain one of each of the 23 chromosomes. In the case of a single chromosome type, it may refer to the case where more or less than two homologous but non-identical chromosome copies are present, or where there are two chromosome copies present that originate from the same parent. In some embodiments, the deletion of a chromosome segment is a microdeletion.

**[73]** *Ploidy State* refers to the quantity and/or chromosomal identity of one or more chromosomes or chromosome segments in a cell.

**[74]** *Chromosome* may refer to a single chromosome copy, meaning a single molecule of DNA of which there are 46 in a normal somatic cell; an example is ‘the maternally derived chromosome 18’. Chromosome may also refer to a chromosome type, of which there are 23 in a normal human somatic cell; an example is ‘chromosome 18’.

**[75]** *Chromosomal Identity* may refer to the referent chromosome number, *i.e.* the chromosome type. Normal humans have 22 types of numbered autosomal chromosome types, and two types of sex chromosomes. It may also refer to the parental origin of the chromosome. It may also refer to a specific chromosome inherited from the parent. It may also refer to other identifying features of a chromosome.

**[76]** *Allelic Data* refers to a set of genotypic data concerning a set of one or more alleles. It may refer to the phased, haplotypic data. It may refer to SNP identities, and it may refer to the sequence data of the DNA, including insertions, deletions, repeats and mutations. It may include the parental origin of each allele.

**[77]** *Allelic State* refers to the actual state of the genes in a set of one or more alleles. It may refer to

the actual state of the genes described by the allelic data.

**[78]** *Allele Count* refers to the number of sequences that map to a particular locus, and if that locus is polymorphic, it refers to the number of sequences that map to each of the alleles. If each allele is counted in a binary fashion, then the allele count will be whole number. If the alleles are counted probabilistically, then the allele count can be a fractional number.

**[79]** *Allele Count Probability* refers to the number of sequences that are likely to map to a particular locus or a set of alleles at a polymorphic locus, combined with the probability of the mapping. Note that allele counts are equivalent to allele count probabilities where the probability of the mapping for each counted sequence is binary (zero or one). In some embodiments, the allele count probabilities may be binary. In some embodiments, the allele count probabilities may be set to be equal to the DNA measurements.

**[80]** *Allelic Distribution*, or “*allele count distribution*” refers to the relative amount of each allele that is present for each locus in a set of loci. An allelic distribution can refer to an individual, to a sample, or to a set of measurements made on a sample. In the context of digital allele measurements such as sequencing, the allelic distribution refers to the number or probable number of reads that map to a particular allele for each allele in a set of polymorphic loci. In the context of analog allele measurements such as SNP arrays, the allelic distribution refers to allele intensities and/or allele ratios. The allele measurements may be treated probabilistically, that is, the likelihood that a given allele is present for a given sequence read is a fraction between 0 and 1, or they may be treated in a binary fashion, that is, any given read is considered to be exactly zero or one copies of a particular allele.

**[81]** *Allelic Distribution Pattern* refers to a set of different allele distributions for different contexts, such as different parental contexts. Certain allelic distribution patterns may be indicative of certain ploidy states.

**[82]** *Allelic Bias* refers to the degree to which the measured ratio of alleles at a heterozygous locus is different to the ratio that was present in the original sample of DNA or RNA. The degree of allelic bias at a particular locus is equal to the observed allelic ratio at that locus, as measured, divided by the ratio of alleles in the original DNA or RNA sample at that locus. Allelic bias may be due to amplification bias, purification bias, or some other phenomenon that affects different alleles differently.

**[83]** *Primer*, also “*PCi? probe*” refers to a single nucleic acid molecule (such as a DNA molecule or a DNA oligomer) or a collection of nucleic acid molecules (such as DNA molecules or DNA oligomers) where the molecules are identical, or nearly so, and wherein the primer contains a region that is designed

the actual state of the genes described by the allelic data.

**[78]** *Allele Count* refers to the number of sequences that map to a particular locus, and if that locus is polymorphic, it refers to the number of sequences that map to each of the alleles. If each allele is counted in a binary fashion, then the allele count will be whole number. If the alleles are counted probabilistically, then the allele count can be a fractional number.

**[79]** *Allele Count Probability* refers to the number of sequences that are likely to map to a particular locus or a set of alleles at a polymorphic locus, combined with the probability of the mapping. Note that allele counts are equivalent to allele count probabilities where the probability of the mapping for each counted sequence is binary (zero or one). In some embodiments, the allele count probabilities may be binary. In some embodiments, the allele count probabilities may be set to be equal to the DNA measurements.

**[80]** *Allelic Distribution*, or “*allele count distribution*” refers to the relative amount of each allele that is present for each locus in a set of loci. An allelic distribution can refer to an individual, to a sample, or to a set of measurements made on a sample. In the context of digital allele measurements such as sequencing, the allelic distribution refers to the number or probable number of reads that map to a particular allele for each allele in a set of polymorphic loci. In the context of analog allele measurements such as SNP arrays, the allelic distribution refers to allele intensities and/or allele ratios. The allele measurements may be treated probabilistically, that is, the likelihood that a given allele is present for a given sequence read is a fraction between 0 and 1, or they may be treated in a binary fashion, that is, any given read is considered to be exactly zero or one copies of a particular allele.

**[81]** *Allelic Distribution Pattern* refers to a set of different allele distributions for different contexts, such as different parental contexts. Certain allelic distribution patterns may be indicative of certain ploidy states.

**[82]** *Allelic Bias* refers to the degree to which the measured ratio of alleles at a heterozygous locus is different to the ratio that was present in the original sample of DNA or RNA. The degree of allelic bias at a particular locus is equal to the observed allelic ratio at that locus, as measured, divided by the ratio of alleles in the original DNA or RNA sample at that locus. Allelic bias may be due to amplification bias, purification bias, or some other phenomenon that affects different alleles differently.

**[83]** *Primer*, also “*PCi? probe*” refers to a single nucleic acid molecule (such as a DNA molecule or a DNA oligomer) or a collection of nucleic acid molecules (such as DNA molecules or DNA oligomers) where the molecules are identical, or nearly so, and wherein the primer contains a region that is designed

to hybridize to a targeted locus (*e.g.*, a targeted polymorphic locus or a non-polymorphic locus) or to a universal priming sequence, and may contain a priming sequence designed to allow PCR amplification. A primer may also contain a molecular barcode. A primer may contain a random region that differs for each individual molecule.

**[84]** *Library of primers* refers to a population of two or more primers. In various embodiments, the library includes at least 100; 200; 500; 750; 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different primers. In various embodiments, the library includes at least 100; 200; 500; 750; 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different primer pairs, wherein each pair of primers includes a forward test primer and a reverse test primer where each pair of test primers hybridize to a target locus. In some embodiments, the library of primers includes at least 100; 200; 500; 750; 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different individual primers that each hybridize to a different target locus, wherein the individual primers are not part of primer pairs. In some embodiments, the library has both (i) primer pairs and (ii) individual primers (such as universal primers) that are not part of primer pairs.

**[85]** *Different primers* refers to non-identical primers.

**[86]** *Different pools* refers to non-identical pools.

**[87]** *Different target loci* refers to non-identical target loci.

**[88]** *Different amplicons* refers to non-identical amplicons.

**[89]** *Hybrid Capture Probe* refers to any nucleic acid sequence, possibly modified, that is generated by various methods such as PCR or direct synthesis and intended to be complementary to one strand of a specific target DNA sequence in a sample. The exogenous hybrid capture probes may be added to a prepared sample and hybridized through a denature-reannealing process to form duplexes of exogenous-endogenous fragments. These duplexes may then be physically separated from the sample by various means.

**[90]** *Sequence Read* refers to data representing a sequence of nucleotide bases that were measured, *e.g.*, using a clonal sequencing method. Clonal sequencing may produce sequence data representing single, or clones, or clusters of one original DNA molecule. A sequence read may also have associated quality score at each base position of the sequence indicating the probability that nucleotide has been called correctly.

**[91]** *Mapping a sequence read* is the process of determining a sequence read's location of origin in

to hybridize to a targeted locus (*e.g.*, a targeted polymorphic locus or a non-polymorphic locus) or to a universal priming sequence, and may contain a priming sequence designed to allow PCR amplification. A primer may also contain a molecular barcode. A primer may contain a random region that differs for each individual molecule.

**[84]** *Library of primers* refers to a population of two or more primers. In various embodiments, the library includes at least 100; 200; 500; 750; 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different primers. In various embodiments, the library includes at least 100; 200; 500; 750; 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different primer pairs, wherein each pair of primers includes a forward test primer and a reverse test primer where each pair of test primers hybridize to a target locus. In some embodiments, the library of primers includes at least 100; 200; 500; 750; 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different individual primers that each hybridize to a different target locus, wherein the individual primers are not part of primer pairs. In some embodiments, the library has both (i) primer pairs and (ii) individual primers (such as universal primers) that are not part of primer pairs.

**[85]** *Different primers* refers to non-identical primers.

**[86]** *Different pools* refers to non-identical pools.

**[87]** *Different target loci* refers to non-identical target loci.

**[88]** *Different amplicons* refers to non-identical amplicons.

**[89]** *Hybrid Capture Probe* refers to any nucleic acid sequence, possibly modified, that is generated by various methods such as PCR or direct synthesis and intended to be complementary to one strand of a specific target DNA sequence in a sample. The exogenous hybrid capture probes may be added to a prepared sample and hybridized through a denature-reannealing process to form duplexes of exogenous-endogenous fragments. These duplexes may then be physically separated from the sample by various means.

**[90]** *Sequence Read* refers to data representing a sequence of nucleotide bases that were measured, *e.g.*, using a clonal sequencing method. Clonal sequencing may produce sequence data representing single, or clones, or clusters of one original DNA molecule. A sequence read may also have associated quality score at each base position of the sequence indicating the probability that nucleotide has been called correctly.

**[91]** *Mapping a sequence read* is the process of determining a sequence read's location of origin in

the genome sequence of a particular organism. The location of origin of sequence reads is based on similarity of nucleotide sequence of the read and the genome sequence.

**[92]** *Matched Copy Error*, also “*Matching Chromosome Aneuploidy*” (MCA), refers to a state of aneuploidy where one cell contains two identical or nearly identical chromosomes. This type of aneuploidy may arise during the formation of the gametes in meiosis, and may be referred to as a meiotic non-disjunction error. This type of error may arise in mitosis. Matching trisomy may refer to the case where three copies of a given chromosome are present in an individual and two of the copies are identical.

**[93]** *Unmatched Copy Error*, also “*Unique Chromosome Aneuploidy* (UCA), refers to a state of aneuploidy where one cell contains two chromosomes that are from the same parent, and that may be homologous but not identical. This type of aneuploidy may arise during meiosis, and may be referred to as a meiotic error. Unmatching trisomy may refer to the case where three copies of a given chromosome are present in an individual and two of the copies are from the same parent, and are homologous, but are not identical. Note that unmatching trisomy may refer to the case where two homologous chromosomes from one parent are present, and where some segments of the chromosomes are identical while other segments are merely homologous.

**[94]** *Homologous Chromosomes* refers to chromosome copies that contain the same set of genes that normally pair up during meiosis.

**[95]** *Identical Chromosomes* refers to chromosome copies that contain the same set of genes, and for each gene they have the same set of alleles that are identical, or nearly identical.

**[96]** *Allele Drop Out (ADO)* refers to the situation where at least one of the base pairs in a set of base pairs from homologous chromosomes at a given allele is not detected.

**[97]** *Locus Drop Out (LDO)* refers to the situation where both base pairs in a set of base pairs from homologous chromosomes at a given allele are not detected.

**[98]** *Homozygous* refers to having similar alleles as corresponding chromosomal loci.

**[99]** *Heterozygous* refers to having dissimilar alleles as corresponding chromosomal loci.

**[100]** *Heterozygosity Rate* refers to the rate of individuals in the population having heterozygous alleles at a given locus. The heterozygosity rate may also refer to the expected or measured ratio of alleles, at a given locus in an individual, or a sample of DNA or RNA.

**[101]** *Chromosomal Region* refers to a segment of a chromosome, or a full chromosome.

**[102]** *Segment of a Chromosome* refers to a section of a chromosome that can range in size from one base pair to the entire chromosome.

the genome sequence of a particular organism. The location of origin of sequence reads is based on similarity of nucleotide sequence of the read and the genome sequence.

**[92]** *Matched Copy Error*, also “*Matching Chromosome Aneuploidy*” (MCA), refers to a state of aneuploidy where one cell contains two identical or nearly identical chromosomes. This type of aneuploidy may arise during the formation of the gametes in meiosis, and may be referred to as a meiotic non-disjunction error. This type of error may arise in mitosis. Matching trisomy may refer to the case where three copies of a given chromosome are present in an individual and two of the copies are identical.

**[93]** *Unmatched Copy Error*, also “*Unique Chromosome Aneuploidy* (UCA), refers to a state of aneuploidy where one cell contains two chromosomes that are from the same parent, and that may be homologous but not identical. This type of aneuploidy may arise during meiosis, and may be referred to as a meiotic error. Unmatching trisomy may refer to the case where three copies of a given chromosome are present in an individual and two of the copies are from the same parent, and are homologous, but are not identical. Note that unmatching trisomy may refer to the case where two homologous chromosomes from one parent are present, and where some segments of the chromosomes are identical while other segments are merely homologous.

**[94]** *Homologous Chromosomes* refers to chromosome copies that contain the same set of genes that normally pair up during meiosis.

**[95]** *Identical Chromosomes* refers to chromosome copies that contain the same set of genes, and for each gene they have the same set of alleles that are identical, or nearly identical.

**[96]** *Allele Drop Out (ADO)* refers to the situation where at least one of the base pairs in a set of base pairs from homologous chromosomes at a given allele is not detected.

**[97]** *Locus Drop Out (LDO)* refers to the situation where both base pairs in a set of base pairs from homologous chromosomes at a given allele are not detected.

**[98]** *Homozygous* refers to having similar alleles as corresponding chromosomal loci.

**[99]** *Heterozygous* refers to having dissimilar alleles as corresponding chromosomal loci.

**[100]** *Heterozygosity Rate* refers to the rate of individuals in the population having heterozygous alleles at a given locus. The heterozygosity rate may also refer to the expected or measured ratio of alleles, at a given locus in an individual, or a sample of DNA or RNA.

**[101]** *Chromosomal Region* refers to a segment of a chromosome, or a full chromosome.

**[102]** *Segment of a Chromosome* refers to a section of a chromosome that can range in size from one base pair to the entire chromosome.

**[103]** *Chromosome* refers to either a full chromosome, or a segment or section of a chromosome.

**[104]** *Copies* refers to the number of copies of a chromosome segment. It may refer to identical copies, or to non-identical, homologous copies of a chromosome segment wherein the different copies of the chromosome segment contain a substantially similar set of loci, and where one or more of the alleles are different. Note that in some cases of aneuploidy, such as the M2 copy error, it is possible to have some copies of the given chromosome segment that are identical as well as some copies of the same chromosome segment that are not identical.

**[105]** *Haplotype* refers to a combination of alleles at multiple loci that are typically inherited together on the same chromosome. Haplotype may refer to as few as two loci or to an entire chromosome depending on the number of recombination events that have occurred between a given set of loci. Haplotype can also refer to a set of SNPs on a single chromatid that are statistically associated.

**[106]** *Haplotypic Data*, also “*Phased Data*” or “*Ordered Genetic Data*,” refers to data from a single chromosome or chromosome segment in a diploid or polyploid genome, *e.g.*, either the segregated maternal or paternal copy of a chromosome in a diploid genome.

**[107]** *Phasing* refers to the act of determining the haplotypic genetic data of an individual given unordered, diploid (or polyploidy) genetic data. It may refer to the act of determining which of two genes at an allele, for a set of alleles found on one chromosome, are associated with each of the two homologous chromosomes in an individual.

**[108]** *Phased Data* refers to genetic data where one or more haplotypes have been determined.

**[109]** *Hypothesis* refers to a possible state, such as a possible degree of overrepresentation of the number of copies of a first homologous chromosome or chromosome segment as compared to a second homologous chromosome or chromosome segment, a possible deletion, a possible duplication, a possible ploidy state at a given set of one or more chromosomes or chromosome segments, a possible allelic state at a given set of one or more loci, a possible paternity relationship, or a possible DNA, RNA, fetal fraction at a given set of one or more chromosomes or chromosome segment, or a set of quantities of genetic material from a set of loci. The genetic states can optionally be linked with probabilities indicating the relative likelihood of each of the elements in the hypothesis being true in relation to other elements in the hypothesis, or the relative likelihood of the hypothesis as a whole being true. The set of possibilities may comprise one or more elements.

**[110]** *Copy Number Hypothesis*, also “*Ploidy State Hypothesis*,” refers to a hypothesis concerning the number of copies of a chromosome or chromosome segment in an individual. It may also refer to a

**[103]** *Chromosome* refers to either a full chromosome, or a segment or section of a chromosome.

**[104]** *Copies* refers to the number of copies of a chromosome segment. It may refer to identical copies, or to non-identical, homologous copies of a chromosome segment wherein the different copies of the chromosome segment contain a substantially similar set of loci, and where one or more of the alleles are different. Note that in some cases of aneuploidy, such as the M2 copy error, it is possible to have some copies of the given chromosome segment that are identical as well as some copies of the same chromosome segment that are not identical.

**[105]** *Haplotype* refers to a combination of alleles at multiple loci that are typically inherited together on the same chromosome. Haplotype may refer to as few as two loci or to an entire chromosome depending on the number of recombination events that have occurred between a given set of loci. Haplotype can also refer to a set of SNPs on a single chromatid that are statistically associated.

**[106]** *Haplotypic Data*, also “*Phased Data*” or “*Ordered Genetic Data*,” refers to data from a single chromosome or chromosome segment in a diploid or polyploid genome, *e.g.*, either the segregated maternal or paternal copy of a chromosome in a diploid genome.

**[107]** *Phasing* refers to the act of determining the haplotypic genetic data of an individual given unordered, diploid (or polyploidy) genetic data. It may refer to the act of determining which of two genes at an allele, for a set of alleles found on one chromosome, are associated with each of the two homologous chromosomes in an individual.

**[108]** *Phased Data* refers to genetic data where one or more haplotypes have been determined.

**[109]** *Hypothesis* refers to a possible state, such as a possible degree of overrepresentation of the number of copies of a first homologous chromosome or chromosome segment as compared to a second homologous chromosome or chromosome segment, a possible deletion, a possible duplication, a possible ploidy state at a given set of one or more chromosomes or chromosome segments, a possible allelic state at a given set of one or more loci, a possible paternity relationship, or a possible DNA, RNA, fetal fraction at a given set of one or more chromosomes or chromosome segment, or a set of quantities of genetic material from a set of loci. The genetic states can optionally be linked with probabilities indicating the relative likelihood of each of the elements in the hypothesis being true in relation to other elements in the hypothesis, or the relative likelihood of the hypothesis as a whole being true. The set of possibilities may comprise one or more elements.

**[110]** *Copy Number Hypothesis*, also “*Ploidy State Hypothesis*,” refers to a hypothesis concerning the number of copies of a chromosome or chromosome segment in an individual. It may also refer to a

hypothesis concerning the identity of each of the chromosomes, including the parent of origin of each chromosome, and which of the parent's two chromosomes are present in the individual. It may also refer to a hypothesis concerning which chromosomes, or chromosome segments, if any, from a related individual correspond genetically to a given chromosome from an individual.

**[111]** *Related Individual* refers to any individual who is genetically related to, and thus shares haplotype blocks with, the target individual. In one context, the related individual may be a genetic parent of the target individual, or any genetic material derived from a parent, such as a sperm, a polar body, an embryo, a fetus, or a child. It may also refer to a sibling, parent, or grandparent.

**[112]** *Sibling* refers to any individual whose genetic parents are the same as the individual in question. In some embodiments, it may refer to a bom child, an embryo, or a fetus, or one or more cells originating from a bom child, an embryo, or a fetus. A sibling may also refer to a haploid individual that originates from one of the parents, such as a sperm, a polar body, or any other set of haplotypic genetic matter. An individual may be considered to be a sibling of itself.

**[113]** *Child* may refer to an embryo, a blastomere, or a fetus. Note that in the presently disclosed embodiments, the concepts described apply equally well to individuals who are a bom child, a fetus, an embryo, or a set of cells therefrom. The use of the term child may simply be meant to connote that the individual referred to as the child is the genetic offspring of the parents.

**[114]** *Fetal* refers to “of the fetus,” or “of the region of the placenta that is genetically similar to the fetus”. In a pregnant woman, some portion of the placenta is genetically similar to the fetus, and the free floating fetal DNA found in maternal blood may have originated from the portion of the placenta with a genotype that matches the fetus. Note that the genetic information in half of the chromosomes in a fetus is inherited from the mother of the fetus. In some embodiments, the DNA from these maternally inherited chromosomes that came from a fetal cell is considered to be “of fetal origin,” not “of maternal origin.”

**[115]** *DNA of Fetal Origin* refers to DNA that was originally part of a cell whose genotype was essentially equivalent to that of the fetus.

**[116]** *DNA of Maternal Origin* refers to DNA that was originally part of a cell whose genotype was essentially equivalent to that of the mother.

**[117]** *Parent* refers to the genetic mother or father of an individual. An individual typically has two parents, a mother and a father, though this may not necessarily be the case such as in genetic or chromosomal chimerism. A parent may be considered to be an individual.

**[118]** *Parental Context* refers to the genetic state of a given SNP, on each of the two relevant

hypothesis concerning the identity of each of the chromosomes, including the parent of origin of each chromosome, and which of the parent's two chromosomes are present in the individual. It may also refer to a hypothesis concerning which chromosomes, or chromosome segments, if any, from a related individual correspond genetically to a given chromosome from an individual.

**[111]** *Related Individual* refers to any individual who is genetically related to, and thus shares haplotype blocks with, the target individual. In one context, the related individual may be a genetic parent of the target individual, or any genetic material derived from a parent, such as a sperm, a polar body, an embryo, a fetus, or a child. It may also refer to a sibling, parent, or grandparent.

**[112]** *Sibling* refers to any individual whose genetic parents are the same as the individual in question. In some embodiments, it may refer to a bom child, an embryo, or a fetus, or one or more cells originating from a bom child, an embryo, or a fetus. A sibling may also refer to a haploid individual that originates from one of the parents, such as a sperm, a polar body, or any other set of haplotypic genetic matter. An individual may be considered to be a sibling of itself.

**[113]** *Child* may refer to an embryo, a blastomere, or a fetus. Note that in the presently disclosed embodiments, the concepts described apply equally well to individuals who are a bom child, a fetus, an embryo, or a set of cells therefrom. The use of the term child may simply be meant to connote that the individual referred to as the child is the genetic offspring of the parents.

**[114]** *Fetal* refers to “of the fetus,” or “of the region of the placenta that is genetically similar to the fetus”. In a pregnant woman, some portion of the placenta is genetically similar to the fetus, and the free floating fetal DNA found in maternal blood may have originated from the portion of the placenta with a genotype that matches the fetus. Note that the genetic information in half of the chromosomes in a fetus is inherited from the mother of the fetus. In some embodiments, the DNA from these maternally inherited chromosomes that came from a fetal cell is considered to be “of fetal origin,” not “of maternal origin.”

**[115]** *DNA of Fetal Origin* refers to DNA that was originally part of a cell whose genotype was essentially equivalent to that of the fetus.

**[116]** *DNA of Maternal Origin* refers to DNA that was originally part of a cell whose genotype was essentially equivalent to that of the mother.

**[117]** *Parent* refers to the genetic mother or father of an individual. An individual typically has two parents, a mother and a father, though this may not necessarily be the case such as in genetic or chromosomal chimerism. A parent may be considered to be an individual.

**[118]** *Parental Context* refers to the genetic state of a given SNP, on each of the two relevant

chromosomes for one or both of the two parents of the target.

**[119]** *Maternal Plasma* refers to the plasma portion of the blood from a female who is pregnant.

**[120]** *Clinical Decision* refers to any decision to take or not take an action that has an outcome that affects the health or survival of an individual. A clinical decision may also refer to a decision to conduct further testing, to abort or maintain a pregnancy, to take actions to mitigate an undesirable phenotype, or to take actions to prepare for a phenotype.

**[121]** *Diagnostic Box* refers to one or a combination of machines designed to perform one or a plurality of aspects of the methods disclosed herein. In an embodiment, the diagnostic box may be placed at a point of patient care. In an embodiment, the diagnostic box may perform targeted amplification followed by sequencing. In an embodiment the diagnostic box may function alone or with the help of a technician.

**[122]** *Informatics Based Method* refers to a method that relies heavily on statistics to make sense of a large amount of data. In the context of prenatal diagnosis, it refers to a method designed to determine the ploidy state at one or more chromosomes or chromosome segments, the allelic state at one or more alleles, or paternity by statistically inferring the most likely state, rather than by directly physically measuring the state, given a large amount of genetic data, for example from a molecular array or sequencing. In an embodiment of the present disclosure, the informatics based technique may be one disclosed in this patent application. In an embodiment of the present disclosure it may be PARENTAL SUPPORT.

**[123]** *Primary Genetic Data* refers to the analog intensity signals that are output by a genotyping platform. In the context of SNP arrays, primary genetic data refers to the intensity signals before any genotype calling has been done. In the context of sequencing, primary genetic data refers to the analog measurements, analogous to the chromatogram, that comes off the sequencer before the identity of any base pairs have been determined, and before the sequence has been mapped to the genome.

**[124]** *Secondary Genetic Data* refers to processed genetic data that are output by a genotyping platform. In the context of a SNP array, the secondary genetic data refers to the allele calls made by software associated with the SNP array reader, wherein the software has made a call whether a given allele is present or not present in the sample. In the context of sequencing, the secondary genetic data refers to the base pair identities of the sequences have been determined, and possibly also where the sequences have been mapped to the genome.

**[125]** *Preferential Enrichment* of DNA that corresponds to a locus, or preferential enrichment of DNA at a locus, refers to any method that results in the percentage of molecules of DNA in a post-enrichment DNA mixture that correspond to the locus being higher than the percentage of molecules of DNA in the

chromosomes for one or both of the two parents of the target.

**[119]** *Maternal Plasma* refers to the plasma portion of the blood from a female who is pregnant.

**[120]** *Clinical Decision* refers to any decision to take or not take an action that has an outcome that affects the health or survival of an individual. A clinical decision may also refer to a decision to conduct further testing, to abort or maintain a pregnancy, to take actions to mitigate an undesirable phenotype, or to take actions to prepare for a phenotype.

**[121]** *Diagnostic Box* refers to one or a combination of machines designed to perform one or a plurality of aspects of the methods disclosed herein. In an embodiment, the diagnostic box may be placed at a point of patient care. In an embodiment, the diagnostic box may perform targeted amplification followed by sequencing. In an embodiment the diagnostic box may function alone or with the help of a technician.

**[122]** *Informatics Based Method* refers to a method that relies heavily on statistics to make sense of a large amount of data. In the context of prenatal diagnosis, it refers to a method designed to determine the ploidy state at one or more chromosomes or chromosome segments, the allelic state at one or more alleles, or paternity by statistically inferring the most likely state, rather than by directly physically measuring the state, given a large amount of genetic data, for example from a molecular array or sequencing. In an embodiment of the present disclosure, the informatics based technique may be one disclosed in this patent application. In an embodiment of the present disclosure it may be PARENTAL SUPPORT.

**[123]** *Primary Genetic Data* refers to the analog intensity signals that are output by a genotyping platform. In the context of SNP arrays, primary genetic data refers to the intensity signals before any genotype calling has been done. In the context of sequencing, primary genetic data refers to the analog measurements, analogous to the chromatogram, that comes off the sequencer before the identity of any base pairs have been determined, and before the sequence has been mapped to the genome.

**[124]** *Secondary Genetic Data* refers to processed genetic data that are output by a genotyping platform. In the context of a SNP array, the secondary genetic data refers to the allele calls made by software associated with the SNP array reader, wherein the software has made a call whether a given allele is present or not present in the sample. In the context of sequencing, the secondary genetic data refers to the base pair identities of the sequences have been determined, and possibly also where the sequences have been mapped to the genome.

**[125]** *Preferential Enrichment* of DNA that corresponds to a locus, or preferential enrichment of DNA at a locus, refers to any method that results in the percentage of molecules of DNA in a post-enrichment DNA mixture that correspond to the locus being higher than the percentage of molecules of DNA in the

pre-enrichment DNA mixture that correspond to the locus. The method may involve selective amplification of DNA molecules that correspond to a locus. The method may involve removing DNA molecules that do not correspond to the locus. The method may involve a combination of methods. The degree of enrichment is defined as the percentage of molecules of DNA in the post-enrichment mixture that correspond to the locus divided by the percentage of molecules of DNA in the pre-enrichment mixture that correspond to the locus. Preferential enrichment may be carried out at a plurality of loci. In some embodiments of the present disclosure, the degree of enrichment is greater than 20, 200, or 2,000. When preferential enrichment is carried out at a plurality of loci, the degree of enrichment may refer to the average degree of enrichment of all of the loci in the set of loci.

**[126]** *Amplification* refers to a method that increases the number of copies of a molecule of DNA or RNA.

**[127]** *Selective Amplification* may refer to a method that increases the number of copies of a particular molecule of DNA (or RNA), or molecules of DNA (or RNA) that correspond to a particular region of DNA (or RNA). It may also refer to a method that increases the number of copies of a particular targeted molecule of DNA (or RNA), or targeted region of DNA (or RNA) more than it increases non-targeted molecules or regions of DNA (or RNA). Selective amplification may be a method of preferential enrichment.

**[128]** *Universal Priming Sequence* refers to a DNA (or RNA) sequence that may be appended to a population of target DNA (or RNA) molecules, for example by ligation, PCR, or ligation mediated PCR. Once added to the population of target molecules, primers specific to the universal priming sequences can be used to amplify the target population using a single pair of amplification primers. Universal priming sequences are typically not related to the target sequences.

**[129]** *Universal Adapters*, or “*ligation adaptors*” or “*library tags*” are nucleic acid molecules containing a universal priming sequence that can be covalently linked to the 5-prime and 3-prime end of a population of target double stranded nucleic acid molecules. The addition of the adapters provides universal priming sequences to the 5-prime and 3-prime end of the target population from which PCR amplification can take place, amplifying all molecules from the target population, using a single pair of amplification primers.

**[130]** *Targeting* refers to a method used to selectively amplify or otherwise preferentially enrich those molecules of DNA (or RNA) that correspond to a set of loci in a mixture of DNA (or RNA).

**[131]** *Joint Distribution Model* refers to a model that defines the probability of events defined in terms

pre-enrichment DNA mixture that correspond to the locus. The method may involve selective amplification of DNA molecules that correspond to a locus. The method may involve removing DNA molecules that do not correspond to the locus. The method may involve a combination of methods. The degree of enrichment is defined as the percentage of molecules of DNA in the post-enrichment mixture that correspond to the locus divided by the percentage of molecules of DNA in the pre-enrichment mixture that correspond to the locus. Preferential enrichment may be carried out at a plurality of loci. In some embodiments of the present disclosure, the degree of enrichment is greater than 20, 200, or 2,000. When preferential enrichment is carried out at a plurality of loci, the degree of enrichment may refer to the average degree of enrichment of all of the loci in the set of loci.

**[126]** *Amplification* refers to a method that increases the number of copies of a molecule of DNA or RNA.

**[127]** *Selective Amplification* may refer to a method that increases the number of copies of a particular molecule of DNA (or RNA), or molecules of DNA (or RNA) that correspond to a particular region of DNA (or RNA). It may also refer to a method that increases the number of copies of a particular targeted molecule of DNA (or RNA), or targeted region of DNA (or RNA) more than it increases non-targeted molecules or regions of DNA (or RNA). Selective amplification may be a method of preferential enrichment.

**[128]** *Universal Priming Sequence* refers to a DNA (or RNA) sequence that may be appended to a population of target DNA (or RNA) molecules, for example by ligation, PCR, or ligation mediated PCR. Once added to the population of target molecules, primers specific to the universal priming sequences can be used to amplify the target population using a single pair of amplification primers. Universal priming sequences are typically not related to the target sequences.

**[129]** *Universal Adapters*, or “*ligation adaptors*” or “*library tags*” are nucleic acid molecules containing a universal priming sequence that can be covalently linked to the 5-prime and 3-prime end of a population of target double stranded nucleic acid molecules. The addition of the adapters provides universal priming sequences to the 5-prime and 3-prime end of the target population from which PCR amplification can take place, amplifying all molecules from the target population, using a single pair of amplification primers.

**[130]** *Targeting* refers to a method used to selectively amplify or otherwise preferentially enrich those molecules of DNA (or RNA) that correspond to a set of loci in a mixture of DNA (or RNA).

**[131]** *Joint Distribution Model* refers to a model that defines the probability of events defined in terms

of multiple random variables, given a plurality of random variables defined on the same probability space, where the probabilities of the variable are linked. In some embodiments, the degenerate case where the probabilities of the variables are not linked may be used.

**[132]** *Cancer-related gene* refers to a gene associated with an altered risk for a cancer or an altered prognosis for a cancer. Exemplary cancer-related genes that promote cancer include oncogenes; genes that enhance cell proliferation, invasion, or metastasis; genes that inhibit apoptosis; and pro-angiogenesis genes. Cancer-related genes that inhibit cancer include, but are not limited to, tumor suppressor genes; genes that inhibit cell proliferation, invasion, or metastasis; genes that promote apoptosis; and anti-angiogenesis genes.

**[133]** *Estrogen-related cancer* refers to a cancer that is modulated by estrogen. Examples of estrogen-related cancers include, without limitation, breast cancer and ovarian cancer. Her2 is overexpressed in many estrogen-related cancers (U.S. Pat. No. 6,165,464, which is hereby incorporated by reference in its entirety).

**[134]** *Androgen-related cancer* refers to a cancer that is modulated by androgen. An example of androgen-related cancers is prostate cancer.

**[135]** *Higher than normal expression level* refers to expression of an mRNA or protein at a level that is higher than the average expression level of the corresponding molecule in control subjects (such as subjects without a disease or disorder such as cancer). In various embodiments, the expression level is at least 20, 40, 50, 75, 90, 100, 200, 500, or even 1000% higher than the level in control subjects.

**[136]** *Lower than normal expression level* refers to expression of an mRNA or protein at a level that is lower than the average expression level of the corresponding molecule in control subjects (such as subjects without a disease or disorder such as cancer). In various embodiments, the expression level is at least 20, 40, 50, 75, 90, 95, or 100% lower than the level in control subjects. In some embodiments, the expression of the mRNA or protein is not detectable.

**[137]** *Modulate expression or activity* refers to either increasing or decreasing expression or activity, for example, of a protein or nucleic acid sequence, relative to control conditions. In some embodiments, the modulation in expression or activity is an increase or decrease of at least 10, 20, 40, 50, 75, 90, 100, 200, 500, or even 1000%). In various embodiments, transcription, translation, mRNA or protein stability, or the binding of the mRNA or protein to other molecules *in vivo* is modulated by the therapy. In some embodiments, the level of mRNA is determined by standard Northern blot analysis, and the level of protein is determined by standard Western blot analysis, such as the analyses described herein or those described

of multiple random variables, given a plurality of random variables defined on the same probability space, where the probabilities of the variable are linked. In some embodiments, the degenerate case where the probabilities of the variables are not linked may be used.

**[132]** *Cancer-related gene* refers to a gene associated with an altered risk for a cancer or an altered prognosis for a cancer. Exemplary cancer-related genes that promote cancer include oncogenes; genes that enhance cell proliferation, invasion, or metastasis; genes that inhibit apoptosis; and pro-angiogenesis genes. Cancer-related genes that inhibit cancer include, but are not limited to, tumor suppressor genes; genes that inhibit cell proliferation, invasion, or metastasis; genes that promote apoptosis; and anti-angiogenesis genes.

**[133]** *Estrogen-related cancer* refers to a cancer that is modulated by estrogen. Examples of estrogen-related cancers include, without limitation, breast cancer and ovarian cancer. Her2 is overexpressed in many estrogen-related cancers (U.S. Pat. No. 6,165,464, which is hereby incorporated by reference in its entirety).

**[134]** *Androgen-related cancer* refers to a cancer that is modulated by androgen. An example of androgen-related cancers is prostate cancer.

**[135]** *Higher than normal expression level* refers to expression of an mRNA or protein at a level that is higher than the average expression level of the corresponding molecule in control subjects (such as subjects without a disease or disorder such as cancer). In various embodiments, the expression level is at least 20, 40, 50, 75, 90, 100, 200, 500, or even 1000% higher than the level in control subjects.

**[136]** *Lower than normal expression level* refers to expression of an mRNA or protein at a level that is lower than the average expression level of the corresponding molecule in control subjects (such as subjects without a disease or disorder such as cancer). In various embodiments, the expression level is at least 20, 40, 50, 75, 90, 95, or 100% lower than the level in control subjects. In some embodiments, the expression of the mRNA or protein is not detectable.

**[137]** *Modulate expression or activity* refers to either increasing or decreasing expression or activity, for example, of a protein or nucleic acid sequence, relative to control conditions. In some embodiments, the modulation in expression or activity is an increase or decrease of at least 10, 20, 40, 50, 75, 90, 100, 200, 500, or even 1000%). In various embodiments, transcription, translation, mRNA or protein stability, or the binding of the mRNA or protein to other molecules *in vivo* is modulated by the therapy. In some embodiments, the level of mRNA is determined by standard Northern blot analysis, and the level of protein is determined by standard Western blot analysis, such as the analyses described herein or those described

by, for example, Ausubel *et al.* (Current Protocols in Molecular Biology, John Wiley & Sons, New York, July 11, 2013, which is hereby incorporated by reference in its entirety). In one embodiment, the level of a protein is determined by measuring the level of enzymatic activity, using standard methods. In another preferred embodiment, the level of mRNA, protein, or enzymatic activity is equal to or less than 20, 10, 5, or 2-fold above the corresponding level in control cells that do not express a functional form of the protein, such as cells homozygous for a nonsense mutation. In yet another embodiment, the level of mRNA, protein, or enzymatic activity is equal to or less than 20, 10, 5, or 2-fold above the corresponding basal level in control cells, such as non-cancerous cells, cells that have not been exposed to conditions that induce abnormal cell proliferation or that inhibit apoptosis, or cells from a subject without the disease or disorder of interest.

**[138]** *Dosage sufficient to modulate mRNA or protein expression or activity* refers to an amount of a therapy that increases or decreases mRNA or protein expression or activity when administered to a subject. In some embodiments, for a compound that decreases expression or activity, the modulation is a decrease in expression or activity that is at least 10%, 30%, 40%, 50%, 75%, or 90% lower in a treated subject than in the same subject prior to the administration of the inhibitor or than in an untreated, control subject. In addition, In some embodiments, for a compound that increases expression or activity, the amount of expression or activity of the mRNA or protein is at least 1.5-, 2-, 3-, 5-, 10-, or 20-fold greater in a treated subject than in the same subject prior to the administration of the modulator or than in an untreated, control subject.

**[139]** In some embodiments, compounds may directly or indirectly modulate the expression or activity of the mRNA or protein. For example, a compound may indirectly modulate the expression or activity of an mRNA or protein of interest by modulating the expression or activity of a molecule (*e.g.*, a nucleic acid, protein, signaling molecule, growth factor, cytokine, or chemokine) that directly or indirectly affects the expression or activity of the mRNA or protein of interest. In some embodiments, the compounds inhibit cell division or induce apoptosis. These compounds in the therapy may include, for example, unpurified or purified proteins, antibodies, synthetic organic molecules, naturally-occurring organic molecules, nucleic acid molecules, and components thereof. The compounds in a combination therapy may be administered simultaneously or sequentially. Exemplary compounds include signal transduction inhibitors.

**[140]** *Purified* refers to being separated from other components that naturally accompany it. Typically, a factor is substantially pure when it is at least 50%, by weight, free from proteins, antibodies, and

by, for example, Ausubel *et al.* (Current Protocols in Molecular Biology, John Wiley & Sons, New York, July 11, 2013, which is hereby incorporated by reference in its entirety). In one embodiment, the level of a protein is determined by measuring the level of enzymatic activity, using standard methods. In another preferred embodiment, the level of mRNA, protein, or enzymatic activity is equal to or less than 20, 10, 5, or 2-fold above the corresponding level in control cells that do not express a functional form of the protein, such as cells homozygous for a nonsense mutation. In yet another embodiment, the level of mRNA, protein, or enzymatic activity is equal to or less than 20, 10, 5, or 2-fold above the corresponding basal level in control cells, such as non-cancerous cells, cells that have not been exposed to conditions that induce abnormal cell proliferation or that inhibit apoptosis, or cells from a subject without the disease or disorder of interest.

**[138]** *Dosage sufficient to modulate mRNA or protein expression or activity* refers to an amount of a therapy that increases or decreases mRNA or protein expression or activity when administered to a subject. In some embodiments, for a compound that decreases expression or activity, the modulation is a decrease in expression or activity that is at least 10%, 30%, 40%, 50%, 75%, or 90% lower in a treated subject than in the same subject prior to the administration of the inhibitor or than in an untreated, control subject. In addition, In some embodiments, for a compound that increases expression or activity, the amount of expression or activity of the mRNA or protein is at least 1.5-, 2-, 3-, 5-, 10-, or 20-fold greater in a treated subject than in the same subject prior to the administration of the modulator or than in an untreated, control subject.

**[139]** In some embodiments, compounds may directly or indirectly modulate the expression or activity of the mRNA or protein. For example, a compound may indirectly modulate the expression or activity of an mRNA or protein of interest by modulating the expression or activity of a molecule (*e.g.*, a nucleic acid, protein, signaling molecule, growth factor, cytokine, or chemokine) that directly or indirectly affects the expression or activity of the mRNA or protein of interest. In some embodiments, the compounds inhibit cell division or induce apoptosis. These compounds in the therapy may include, for example, unpurified or purified proteins, antibodies, synthetic organic molecules, naturally-occurring organic molecules, nucleic acid molecules, and components thereof. The compounds in a combination therapy may be administered simultaneously or sequentially. Exemplary compounds include signal transduction inhibitors.

**[140]** *Purified* refers to being separated from other components that naturally accompany it. Typically, a factor is substantially pure when it is at least 50%, by weight, free from proteins, antibodies, and

naturally-occurring organic molecules with which it is naturally associated. In some embodiments, the factor is at least 75%, 90%, or 99%, by weight, pure. A substantially pure factor may be obtained by chemical synthesis, separation of the factor from natural sources, or production of the factor in a recombinant host cell that does not naturally produce the factor.

Proteins and small molecules may be purified by one skilled in the art using standard techniques such as those described by Ausubel *et al.* (Current Protocols in Molecular Biology, John Wiley & Sons, New York, July 11, 2013, which is hereby incorporated by reference in its entirety). In some embodiments the factor is at least 2, 5, or 10 times as pure as the starting material, as measured using polyacrylamide gel electrophoresis, column chromatography, optical density, HPLC analysis, or western analysis (Ausubel *et al.*, *supra*). Exemplary methods of purification include immunoprecipitation, column chromatography such as immunoaffinity chromatography, magnetic bead immunoaffinity purification, and panning with a plate-bound antibody.

[141] Other features and advantages of the invention will be apparent from the following detailed description and from the claims.

#### **BRIEF DESCRIPTION OF THE DRAWINGS**

[142] The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

[143] The presently disclosed embodiments will be further explained with reference to the attached drawings, wherein like structures are referred to by like numerals throughout the several views. The drawings shown are not necessarily to scale, with emphasis instead generally being placed upon illustrating the principles of the presently disclosed embodiments.

[144] FIGs. 1A-1D are graphs showing the distribution of the test statistic  $S$  divided by  $T$  (the number of SNPs) (" $S/T$ ") for various copy number hypotheses for a depth of read (DOR) of 500 and a tumor fraction of 1% for an increasing number of SNPs.

[145] FIGs. 2A-2D are graphs showing the distribution of  $S/T$  for various copy number hypotheses for a DOR of 500 and tumor fraction of 2% for an increasing number of SNPs.

[146] FIGs. 3A-3D are graphs showing the distribution of  $S/T$  for various copy number hypotheses for a DOR of 500 and tumor fraction of 3% for an increasing number of SNPs.

[147] FIGs. 4A-4D are graphs showing the distribution of  $S/T$  for various copy number hypotheses for

naturally-occurring organic molecules with which it is naturally associated. In some embodiments, the factor is at least 75%, 90%, or 99%, by weight, pure. A substantially pure factor may be obtained by chemical synthesis, separation of the factor from natural sources, or production of the factor in a recombinant host cell that does not naturally produce the factor.

Proteins and small molecules may be purified by one skilled in the art using standard techniques such as those described by Ausubel *et al.* (Current Protocols in Molecular Biology, John Wiley & Sons, New York, July 11, 2013, which is hereby incorporated by reference in its entirety). In some embodiments the factor is at least 2, 5, or 10 times as pure as the starting material, as measured using polyacrylamide gel electrophoresis, column chromatography, optical density, HPLC analysis, or western analysis (Ausubel *et al.*, *supra*). Exemplary methods of purification include immunoprecipitation, column chromatography such as immunoaffinity chromatography, magnetic bead immunoaffinity purification, and panning with a plate-bound antibody.

[141] Other features and advantages of the invention will be apparent from the following detailed description and from the claims.

#### **BRIEF DESCRIPTION OF THE DRAWINGS**

[142] The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

[143] The presently disclosed embodiments will be further explained with reference to the attached drawings, wherein like structures are referred to by like numerals throughout the several views. The drawings shown are not necessarily to scale, with emphasis instead generally being placed upon illustrating the principles of the presently disclosed embodiments.

[144] FIGs. 1A-1D are graphs showing the distribution of the test statistic  $S$  divided by  $T$  (the number of SNPs) (" $S/T$ ") for various copy number hypotheses for a depth of read (DOR) of 500 and a tumor fraction of 1% for an increasing number of SNPs.

[145] FIGs. 2A-2D are graphs showing the distribution of  $S/T$  for various copy number hypotheses for a DOR of 500 and tumor fraction of 2% for an increasing number of SNPs.

[146] FIGs. 3A-3D are graphs showing the distribution of  $S/T$  for various copy number hypotheses for a DOR of 500 and tumor fraction of 3% for an increasing number of SNPs.

[147] FIGs. 4A-4D are graphs showing the distribution of  $S/T$  for various copy number hypotheses for

a DOR of 500 and tumor fraction of 4% for an increasing number of SNPs.

**[148]** FIGs. 5A-5D are graphs showing the distribution of S/T for various copy number hypotheses for a DOR of 500 and tumor fraction of 5% for an increasing number of SNPs.

**[149]** FIGs. 6A-6D are graphs showing the distribution of S/T for various copy number hypotheses for a DOR of 500 and tumor fraction of 6% for an increasing number of SNPs.

**[150]** FIGs. 7A-7D are graphs showing the distribution of S/T for various copy number hypotheses for a DOR of 1000 and tumor fraction of 0.5% for an increasing number of SNPs.

**[151]** FIGs. 8A-8D are graphs showing the distribution of S/T for various copy number hypotheses for a DOR of 1000 and tumor fraction of 1% for an increasing number of SNPs.

**[152]** FIGs. 9A-9D are graphs showing the distribution of S/T for various copy number hypotheses for a DOR of 1000 and tumor fraction of 2% for an increasing number of SNPs.

**[153]** FIGs. 10A-10D are graphs showing the distribution of S/T for various copy number hypotheses for a DOR of 1000 and tumor fraction of 3% for an increasing number of SNPs.

**[154]** FIGs. 11A-11D are graphs showing the distribution of S/T for various copy number hypotheses for a DOR of 1000 and tumor fraction of 4% for an increasing number of SNPs.

**[155]** FIGs. 12A-12D are graphs showing the distribution of S/T for various copy number hypotheses for a DOR of 3000 and tumor fraction of 0.5% for an increasing number of SNPs.

**[156]** FIGs. 13A-13D are graphs showing the distribution of S/T for various copy number hypotheses for a DOR of 3000 and tumor fraction of 1% for an increasing number of SNPs.

**[157]** FIG. 14 is a table indicating the sensitivity and specificity for detecting six microdeletion syndromes.

**[158]** FIGs. 15A-15C are graphical representations of euploidy. The x-axis represents the linear position of the individual polymorphic loci along the chromosome, and the y-axis represents the number of A allele reads as a fraction of the total (A+B) allele reads. Maternal and fetal genotypes are indicated to the right of the plots. The plots are color-coded according to maternal genotype, such that red indicates a maternal genotype of AA, blue indicates a maternal genotype of BB, and green indicates a maternal genotype of AB. FIG. 15A is a plot of when two chromosomes are present, and the fetal cfDNA fraction is 0%. This plot is from a non-pregnant woman, and thus represents the pattern when the genotype is entirely maternal. Allele clusters are thus centered around 1 (AA alleles), 0.5 (AB alleles), and 0 (BB alleles). FIG. 15B is a plot of when two chromosomes are present, and the fetal fraction is 12%. The contribution of fetal alleles to the fraction of A allele reads shifts the position of some allele spots up or

a DOR of 500 and tumor fraction of 4% for an increasing number of SNPs.

**[148]** FIGs. 5A-5D are graphs showing the distribution of S/T for various copy number hypotheses for a DOR of 500 and tumor fraction of 5% for an increasing number of SNPs.

**[149]** FIGs. 6A-6D are graphs showing the distribution of S/T for various copy number hypotheses for a DOR of 500 and tumor fraction of 6% for an increasing number of SNPs.

**[150]** FIGs. 7A-7D are graphs showing the distribution of S/T for various copy number hypotheses for a DOR of 1000 and tumor fraction of 0.5% for an increasing number of SNPs.

**[151]** FIGs. 8A-8D are graphs showing the distribution of S/T for various copy number hypotheses for a DOR of 1000 and tumor fraction of 1% for an increasing number of SNPs.

**[152]** FIGs. 9A-9D are graphs showing the distribution of S/T for various copy number hypotheses for a DOR of 1000 and tumor fraction of 2% for an increasing number of SNPs.

**[153]** FIGs. 10A-10D are graphs showing the distribution of S/T for various copy number hypotheses for a DOR of 1000 and tumor fraction of 3% for an increasing number of SNPs.

**[154]** FIGs. 11A-11D are graphs showing the distribution of S/T for various copy number hypotheses for a DOR of 1000 and tumor fraction of 4% for an increasing number of SNPs.

**[155]** FIGs. 12A-12D are graphs showing the distribution of S/T for various copy number hypotheses for a DOR of 3000 and tumor fraction of 0.5% for an increasing number of SNPs.

**[156]** FIGs. 13A-13D are graphs showing the distribution of S/T for various copy number hypotheses for a DOR of 3000 and tumor fraction of 1% for an increasing number of SNPs.

**[157]** FIG. 14 is a table indicating the sensitivity and specificity for detecting six microdeletion syndromes.

**[158]** FIGs. 15A-15C are graphical representations of euploidy. The x-axis represents the linear position of the individual polymorphic loci along the chromosome, and the y-axis represents the number of A allele reads as a fraction of the total (A+B) allele reads. Maternal and fetal genotypes are indicated to the right of the plots. The plots are color-coded according to maternal genotype, such that red indicates a maternal genotype of AA, blue indicates a maternal genotype of BB, and green indicates a maternal genotype of AB. FIG. 15A is a plot of when two chromosomes are present, and the fetal cfDNA fraction is 0%. This plot is from a non-pregnant woman, and thus represents the pattern when the genotype is entirely maternal. Allele clusters are thus centered around 1 (AA alleles), 0.5 (AB alleles), and 0 (BB alleles). FIG. 15B is a plot of when two chromosomes are present, and the fetal fraction is 12%. The contribution of fetal alleles to the fraction of A allele reads shifts the position of some allele spots up or

down along the y-axis. FIG. 15C is a plot of when two chromosomes are present, and the fetal fraction is 26%. The pattern, including two red and two blue peripheral bands and a trio of central green bands, is readily apparent.

**[159]** FIGs. 16A and 16B are graphical representations of 22q11.2 deletion syndrome. FIG. 16A is for maternal 22q11.2 deletion carrier (as indicated by the absence of the green AB SNPs). FIG 16B is for a paternally inherited 22q11 deletion in a fetus (as indicated by the presence of one red and one blue peripheral band). The x-axis represents the linear position of the SNPs, and the y-axis indicates the fraction of A allele reads out of the total reads. Each spot represents a single SNP locus.

**[160]** FIG. 17 is a graphical representation of maternally inherited Cri-du-Chat deletion syndrome (as indicated by the presence of two central green bands instead of three green bands). The x-axis represents the linear position of the SNPs, and the y-axis indicates the fraction of A allele reads out of the total reads. Each spot represents a single SNP locus.

**[161]** FIG. 18 is a graphical representation of paternally inherited Wolf-Hirschhorn deletion syndrome (as indicated by the presence of one red and one blue peripheral band). The x-axis represents the linear position of the SNPs, and the y-axis indicates the fraction of A allele reads out of the total reads. Each spot represents a single SNP locus.

**[162]** FIGs. 19A-19D are graphical representations of X chromosome spike-in experiments to represent an extra copy of a chromosome or chromosome segment. The plots show different amounts of DNA from a father mixed with DNA from the daughter: 16% father DNA (FIG. 19A), 10% father DNA (FIG. 19B), 1% father DNA (FIG. 19C), and 0.1% father DNA (FIG. 19D). The x-axis represents the linear position of the SNPs on the X chromosome, and the y-axis indicates the fraction of M allele reads out of the total reads (M + R). Each spot represents a single SNP locus with allele M or R.

**[163]** FIGs. 20A and 20B are graphs of the false negative rate using haplotype data (FIG. 20A) and without haplotype data (FIG. 20B).

**[164]** FIGs. 21A and 21B are graphs of the false positive rate for  $p = 1\%$  using haplotype data (FIG. 21A) and without haplotype data (FIG. 21B).

**[165]** FIGs. 22A and 22B are graphs of the false positive rate for  $p = 1.5\%$  using haplotype data (FIG. 22A) and without haplotype data (FIG. 22B).

**[166]** FIGs. 23A and 23B are graphs of the false positive rate for  $p = 2\%$  using haplotype data (FIG. 23A) and without haplotype data (FIG. 23B).

**[167]** FIGs. 24A and 24B are graphs of the false positive rate for  $p = 2.5\%$  using haplotype data (FIG.

down along the y-axis. FIG. 15C is a plot of when two chromosomes are present, and the fetal fraction is 26%. The pattern, including two red and two blue peripheral bands and a trio of central green bands, is readily apparent.

**[159]** FIGs. 16A and 16B are graphical representations of 22q11.2 deletion syndrome. FIG. 16A is for maternal 22q11.2 deletion carrier (as indicated by the absence of the green AB SNPs). FIG 16B is for a paternally inherited 22q11 deletion in a fetus (as indicated by the presence of one red and one blue peripheral band). The x-axis represents the linear position of the SNPs, and the y-axis indicates the fraction of A allele reads out of the total reads. Each spot represents a single SNP locus.

**[160]** FIG. 17 is a graphical representation of maternally inherited Cri-du-Chat deletion syndrome (as indicated by the presence of two central green bands instead of three green bands). The x-axis represents the linear position of the SNPs, and the y-axis indicates the fraction of A allele reads out of the total reads. Each spot represents a single SNP locus.

**[161]** FIG. 18 is a graphical representation of paternally inherited Wolf-Hirschhorn deletion syndrome (as indicated by the presence of one red and one blue peripheral band). The x-axis represents the linear position of the SNPs, and the y-axis indicates the fraction of A allele reads out of the total reads. Each spot represents a single SNP locus.

**[162]** FIGs. 19A-19D are graphical representations of X chromosome spike-in experiments to represent an extra copy of a chromosome or chromosome segment. The plots show different amounts of DNA from a father mixed with DNA from the daughter: 16% father DNA (FIG. 19A), 10% father DNA (FIG. 19B), 1% father DNA (FIG. 19C), and 0.1% father DNA (FIG. 19D). The x-axis represents the linear position of the SNPs on the X chromosome, and the y-axis indicates the fraction of M allele reads out of the total reads (M + R). Each spot represents a single SNP locus with allele M or R.

**[163]** FIGs. 20A and 20B are graphs of the false negative rate using haplotype data (FIG. 20A) and without haplotype data (FIG. 20B).

**[164]** FIGs. 21A and 21B are graphs of the false positive rate for  $p = 1\%$  using haplotype data (FIG. 21A) and without haplotype data (FIG. 21B).

**[165]** FIGs. 22A and 22B are graphs of the false positive rate for  $p = 1.5\%$  using haplotype data (FIG. 22A) and without haplotype data (FIG. 22B).

**[166]** FIGs. 23A and 23B are graphs of the false positive rate for  $p = 2\%$  using haplotype data (FIG. 23A) and without haplotype data (FIG. 23B).

**[167]** FIGs. 24A and 24B are graphs of the false positive rate for  $p = 2.5\%$  using haplotype data (FIG.

24 A) and without haplotype data (FIG. 24B).

**[168]** FIGs. 25A and 25B are graphs of the false positive rate for  $p = 3\%$  using haplotype data (FIG. 25A) and without haplotype data (FIG. 25B).

**[169]** FIG. 26 is a table of false positive rates for the first simulation.

**[170]** FIG. 27 is a table of false negative rates for the first simulation.

**[171]** FIG. 28A is a graph of reference counts (counts of one allele, such as the “A” allele) divided by total counts for that locus for a normal (noncancerous) cell line.

**[172]** FIG. 28B is a graph of reference counts divided by total counts for a cancer cell line with a deletion.

**[173]** FIG. 28C is a graph of reference counts divided by total counts for a mixture of DNA from the normal cell line and the cancer cell line.

**[174]** FIG. 29 is a graph of reference counts divided by total counts for a plasma sample from a patient with stage Ha breast cancer with a tumor fraction estimated to be 4.33% (in which 4.33% of the DNA is from tumor cells). The green portion of the graph represents a region in which no CNV is present. The portion of the graph with blue and red represents a region in which a CNV is present and there is a visible separation of the measured allele ratios from the expected allele ratio of 0.5. The blue coloring indicates one haplotype, and the red coloring indicates the other haplotype. Approximately 636 heterozygous SNPs were analyzed in the region of the CNV.

**[175]** FIG. 30 is a graph of reference counts divided by total counts for a plasma sample from a patient with stage lib breast cancer with a tumor fraction estimated to be 0.58%. The green portion of the graph represents a region in which no CNV is present. The portion of the graph with blue and red represents a region in which a CNV is present but there is no clearly visible separation of the measured allele ratios from the expected allele ratio of 0.5. For this analysis, 86 heterozygous SNPs were analyzed in the region of the CNV.

**[176]** FIGs. 31A and 3 IB are graphs showing the maximum likelihood estimation of the tumor fraction. The maximum likelihood estimate is indicated by the peak of the graph and is 4.33% for FIG. 31A and 0.58% for FIG. 3 IB.

**[177]** FIG. 32A is a comparison of the graphs of the log of the odds ratio for various possible tumor fractions for the high tumor fraction sample (4.33%) and the low tumor fraction sample (0.58%). If the log odds ratio is less than 0, the euploid hypothesis is more likely. If the log odds ratio is greater than 0, the presence of a CNV is more likely.

24 A) and without haplotype data (FIG. 24B).

**[168]** FIGs. 25A and 25B are graphs of the false positive rate for  $p = 3\%$  using haplotype data (FIG. 25A) and without haplotype data (FIG. 25B).

**[169]** FIG. 26 is a table of false positive rates for the first simulation.

**[170]** FIG. 27 is a table of false negative rates for the first simulation.

**[171]** FIG. 28A is a graph of reference counts (counts of one allele, such as the “A” allele) divided by total counts for that locus for a normal (noncancerous) cell line.

**[172]** FIG. 28B is a graph of reference counts divided by total counts for a cancer cell line with a deletion.

**[173]** FIG. 28C is a graph of reference counts divided by total counts for a mixture of DNA from the normal cell line and the cancer cell line.

**[174]** FIG. 29 is a graph of reference counts divided by total counts for a plasma sample from a patient with stage Ha breast cancer with a tumor fraction estimated to be 4.33% (in which 4.33% of the DNA is from tumor cells). The green portion of the graph represents a region in which no CNV is present. The portion of the graph with blue and red represents a region in which a CNV is present and there is a visible separation of the measured allele ratios from the expected allele ratio of 0.5. The blue coloring indicates one haplotype, and the red coloring indicates the other haplotype. Approximately 636 heterozygous SNPs were analyzed in the region of the CNV.

**[175]** FIG. 30 is a graph of reference counts divided by total counts for a plasma sample from a patient with stage lib breast cancer with a tumor fraction estimated to be 0.58%. The green portion of the graph represents a region in which no CNV is present. The portion of the graph with blue and red represents a region in which a CNV is present but there is no clearly visible separation of the measured allele ratios from the expected allele ratio of 0.5. For this analysis, 86 heterozygous SNPs were analyzed in the region of the CNV.

**[176]** FIGs. 31A and 3 IB are graphs showing the maximum likelihood estimation of the tumor fraction. The maximum likelihood estimate is indicated by the peak of the graph and is 4.33% for FIG. 31A and 0.58% for FIG. 3 IB.

**[177]** FIG. 32A is a comparison of the graphs of the log of the odds ratio for various possible tumor fractions for the high tumor fraction sample (4.33%) and the low tumor fraction sample (0.58%). If the log odds ratio is less than 0, the euploid hypothesis is more likely. If the log odds ratio is greater than 0, the presence of a CNV is more likely.

**[178]** FIG. 32B is a graph of the probability of a deletion divided by the probability of no deletion for various possible tumor fractions for the low tumor fraction sample (0.58%).

**[179]** FIG. 33 is a graph of the log of the odds ratio for various possible tumor fractions for the low tumor fraction sample (0.58%). FIG. 33 is an enlarged version of the graph in FIG. 32A for the low tumor fraction sample.

**[180]** FIG. 34 is a graph showing the limit of detection for single nucleotide variants in a tumor biopsy using three different methods described in Example 6.

**[181]** FIG. 35 is a graph showing the limit of detection for single nucleotide variants in a plasma sample using three different methods described in Example 6.

**[182]** FIGs. 36A and 36B are graphs of the analysis of genomic DNA (FIG. 36A) or DNA from a single cell (FIG. 36B) using a library of approximately 28,000 primers designed to detect CNVs. The presence of two central bands instead of one central band indicates the presence of a CNV. The x-axis represents the linear position of the SNPs, and the y-axis indicates the fraction of A allele reads out of the total reads.

**[183]** FIGs. 37A and 37B are graphs of the analysis of genomic DNA (FIG. 37A) or DNA from a single cell (FIG. 37B) using a library of approximately 3,000 primers designed to detect CNVs. The presence of two central bands instead of one central band indicates the presence of a CNV. The x-axis represents the linear position of the SNPs, and the y-axis indicates the fraction of A allele reads out of the total reads.

**[184]** FIG. 38 is a graph illustrating the uniformity in DOR for these ~3,000 loci.

**[185]** FIG. 39 is a table comparing error call metrics for genomic DNA and DNA from a single cell.

**[186]** FIG. 40 is a graph of error rates for transition mutations and transversion mutations.

**[187]** FIGs. 41a-d are graphs of Sensitivity of CoNVERGe determined with PlasmArts. **(a)** Correlation between CoNVERGe-calculated AAI and actual input fraction in PlasmArt samples with DNA from a 22q11.2 deletion and matched normal cell lines. **(b)** Correlation between calculated AAI and actual tumour DNA input in PlasmArt samples with DNA from HCC2218 breast cancer cells with chromosome 2p and 2q CNVs and matched normal HCC2218BL cells, containing 0–9.09% tumour DNA fractions. **(c)** Correlation between calculated AAI and actual tumour DNA input in PlasmArt samples with DNA from HCC1954 breast cancer cells with chromosome 1p and 1q CNVs and matched normal HCC1954BL cells, containing 0–5.66% tumour DNA fractions. **(d)** Allele frequency plot for HCC1954 cells used in **(c)**. In **(a)**, **(b)**, and **(c)**, data points and error bars indicate the mean and standard deviation (SD), respectively, of

**[178]** FIG. 32B is a graph of the probability of a deletion divided by the probability of no deletion for various possible tumor fractions for the low tumor fraction sample (0.58%).

**[179]** FIG. 33 is a graph of the log of the odds ratio for various possible tumor fractions for the low tumor fraction sample (0.58%). FIG. 33 is an enlarged version of the graph in FIG. 32A for the low tumor fraction sample.

**[180]** FIG. 34 is a graph showing the limit of detection for single nucleotide variants in a tumor biopsy using three different methods described in Example 6.

**[181]** FIG. 35 is a graph showing the limit of detection for single nucleotide variants in a plasma sample using three different methods described in Example 6.

**[182]** FIGs. 36A and 36B are graphs of the analysis of genomic DNA (FIG. 36A) or DNA from a single cell (FIG. 36B) using a library of approximately 28,000 primers designed to detect CNVs. The presence of two central bands instead of one central band indicates the presence of a CNV. The x-axis represents the linear position of the SNPs, and the y-axis indicates the fraction of A allele reads out of the total reads.

**[183]** FIGs. 37A and 37B are graphs of the analysis of genomic DNA (FIG. 37A) or DNA from a single cell (FIG. 37B) using a library of approximately 3,000 primers designed to detect CNVs. The presence of two central bands instead of one central band indicates the presence of a CNV. The x-axis represents the linear position of the SNPs, and the y-axis indicates the fraction of A allele reads out of the total reads.

**[184]** FIG. 38 is a graph illustrating the uniformity in DOR for these ~3,000 loci.

**[185]** FIG. 39 is a table comparing error call metrics for genomic DNA and DNA from a single cell.

**[186]** FIG. 40 is a graph of error rates for transition mutations and transversion mutations.

**[187]** FIGs. 41a-d are graphs of Sensitivity of CoNVERGe determined with PlasmArts. **(a)** Correlation between CoNVERGe-calculated AAI and actual input fraction in PlasmArt samples with DNA from a 22q11.2 deletion and matched normal cell lines. **(b)** Correlation between calculated AAI and actual tumour DNA input in PlasmArt samples with DNA from HCC2218 breast cancer cells with chromosome 2p and 2q CNVs and matched normal HCC2218BL cells, containing 0–9.09% tumour DNA fractions. **(c)** Correlation between calculated AAI and actual tumour DNA input in PlasmArt samples with DNA from HCC1954 breast cancer cells with chromosome 1p and 1q CNVs and matched normal HCC1954BL cells, containing 0–5.66% tumour DNA fractions. **(d)** Allele frequency plot for HCC1954 cells used in **(c)**. In **(a)**, **(b)**, and **(c)**, data points and error bars indicate the mean and standard deviation (SD), respectively, of

3–8 replicates.

**[188]** FIG. 42 provides details regarding an exemplary Plasmart standard include graphs of fragment size distributions in the lower portion.

**[189]** FIG. 43 right provides results from a dilution curve of Plasmart synthetic ctDNA standards for validation of microdeletion and cancer panels. Right panel shows the maximum likelihood of tumor, estimate of DNA fraction results as an odds ratio plot.

**[190]** FIG. 44 is a plot showing CNVs for various chromosomal regions as indicated for various samples at different % ctDNAs.

**[191]** FIG. 45 is a plot showing CNVs for various chromosomal regions for various ovarian cancer samples with different % ctDNA levels.

**[192]** FIG. 46 is a table showing the percent of breast or lung cancer patients with an SNV or a combined SNV and/or CNV in ctDNA.

**[193]** FIG. 47\_ is a graph of % samples at different breast cancer stages with tumor-specific SNVs and/or CNVs in plasma, and the associated table of data on the right.

**[194]** FIG. 48 is a graph of % samples at different breast cancer substages with tumor-specific SNVs and/or CNVs in plasma, and the associated table of data on the right.

**[195]** FIG. 49\_ is a graph of % samples at different lung cancer stages with tumor-specific SNVs and/or CNVs in plasma, and the associated table of data on the right.

**[196]** FIG. 50\_ is a graph of % samples at different lung cancer substages with tumor-specific SNVs and/or CNVs in plasma, and the associated table of data on the right.

**[197]** Fig. 51A represents the histological finding/history for primary lung tumors analyzed for clonal and subclonal tumor heterogeneity. Fig. 51B is a table of the VAF identities of the biopsied lung tumors by whole genome sequencing and assaying by AmpliSEQ.

**[198]** Fig. 52 illustrates the use of ctDNA from plasma to identify both clonal and subclonal SNA mutations to overcome tumor heterogeneity.

**[199]** Fig. 53 is a table comparing VAF calls by AmpliSeq and mmPCR-NGS for detection of SNVs in primary tumor that were missed by AmpliSeq and SNV mutations identified in ctDNA from plasma.

**[200]** Fig. 54A is a plot of % VAF in Primary Lung Tumor. Fig. 54B is a linear regression plot of AmpliSeq VAF vs. Natera VAF.

**[201]** FIG. 55 is a graph of Pool 1/4 of an 84-plex SNV PCR primer reaction when primer concentration is limited.

3–8 replicates.

**[188]** FIG. 42 provides details regarding an exemplary Plasmart standard include graphs of fragment size distributions in the lower portion.

**[189]** FIG. 43 right provides results from a dilution curve of Plasmart synthetic ctDNA standards for validation of microdeletion and cancer panels. Right panel shows the maximum likelihood of tumor, estimate of DNA fraction results as an odds ratio plot.

**[190]** FIG. 44 is a plot showing CNVs for various chromosomal regions as indicated for various samples at different % ctDNAs.

**[191]** FIG. 45 is a plot showing CNVs for various chromosomal regions for various ovarian cancer samples with different % ctDNA levels.

**[192]** FIG. 46 is a table showing the percent of breast or lung cancer patients with an SNV or a combined SNV and/or CNV in ctDNA.

**[193]** FIG. 47\_ is a graph of % samples at different breast cancer stages with tumor-specific SNVs and/or CNVs in plasma, and the associated table of data on the right.

**[194]** FIG. 48 is a graph of % samples at different breast cancer substages with tumor-specific SNVs and/or CNVs in plasma, and the associated table of data on the right.

**[195]** FIG. 49\_ is a graph of % samples at different lung cancer stages with tumor-specific SNVs and/or CNVs in plasma, and the associated table of data on the right.

**[196]** FIG. 50\_ is a graph of % samples at different lung cancer substages with tumor-specific SNVs and/or CNVs in plasma, and the associated table of data on the right.

**[197]** Fig. 51A represents the histological finding/history for primary lung tumors analyzed for clonal and subclonal tumor heterogeneity. Fig. 51B is a table of the VAF identities of the biopsied lung tumors by whole genome sequencing and assaying by AmpliSEQ.

**[198]** Fig. 52 illustrates the use of ctDNA from plasma to identify both clonal and subclonal SNA mutations to overcome tumor heterogeneity.

**[199]** Fig. 53 is a table comparing VAF calls by AmpliSeq and mmPCR-NGS for detection of SNVs in primary tumor that were missed by AmpliSeq and SNV mutations identified in ctDNA from plasma.

**[200]** Fig. 54A is a plot of % VAF in Primary Lung Tumor. Fig. 54B is a linear regression plot of AmpliSeq VAF vs. Natera VAF.

**[201]** FIG. 55 is a graph of Pool 1/4 of an 84-plex SNV PCR primer reaction when primer concentration is limited.

[202] FIG. 56 is a graph of Pool 2/4 of an 84-plex SNV PCR primer reaction when primer concentration is limited.

[203] FIG. 57 is a graph of Pool 3/4 of an 84-plex SNV PCR primer reaction when primer concentration is limited.

[204] FIG. 58 is a graph of Pool 4/4 of an 84-plex SNV PCR primer reaction when primer concentration is limited.

[205] FIG. 59 illustrates a plot of Limit of Detection (LOD) vs. Depth of Read (DOR) for detection of SNV Transition and Transversion mutations in a 84-plex PCR reaction at 15 PCR cycles.

[206] FIG. 60 illustrates a plot of Limit of Detection (LOD) vs. Depth of Read (DOR) for detection of SNV Transition and Transversion mutations in a 84-plex PCR reaction at 20 PCR cycles.

[207] FIG. 61 illustrates a plot of Limit of Detection (LOD) vs. Depth of Read (DOR) for detection of SNV Transition and Transversion mutations in a 84-plex PCR reaction at 25 PCR cycles.

[208] FIG. 62 is a plot illustrating detection of tumor mosaicism detected in single cells from the tumor.

[201] While the above-identified drawings set forth presently disclosed embodiments, other embodiments are also contemplated, as noted in the discussion. This disclosure presents illustrative embodiments by way of representation and not limitation. Numerous other modifications and embodiments can be devised by those skilled in the art which fall within the scope and spirit of the principles of the presently disclosed embodiments.

#### **DETAILED DESCRIPTION OF THE INVENTION**

[202] The present invention generally relates, at least in part, to improved methods of determining the presence or absence of copy number variations, such as deletions or duplications of chromosome segments or entire chromosomes. The methods are particularly useful for detecting small deletions or duplications, which can be difficult to detect with high specificity and sensitivity using prior methods due to the small amount of data available from the relevant chromosome segment. Methods of the invention can also be used to detect deletions or duplications that are only present in a small percentage of the cells or nucleic acid molecules that are tested. This allows deletions or duplications to be detected prior to the occurrence of disease (such as at a precancerous stage) or in the early stages of disease, such as before a large number of diseased cells (such as cancer cells) with the deletion or duplication accumulate. The more accurate detection of deletions or duplications associated with a disease or disorder enable improved methods for

[202] FIG. 56 is a graph of Pool 2/4 of an 84-plex SNV PCR primer reaction when primer concentration is limited.

[203] FIG. 57 is a graph of Pool 3/4 of an 84-plex SNV PCR primer reaction when primer concentration is limited.

[204] FIG. 58 is a graph of Pool 4/4 of an 84-plex SNV PCR primer reaction when primer concentration is limited.

[205] FIG. 59 illustrates a plot of Limit of Detection (LOD) vs. Depth of Read (DOR) for detection of SNV Transition and Transversion mutations in a 84-plex PCR reaction at 15 PCR cycles.

[206] FIG. 60 illustrates a plot of Limit of Detection (LOD) vs. Depth of Read (DOR) for detection of SNV Transition and Transversion mutations in a 84-plex PCR reaction at 20 PCR cycles.

[207] FIG. 61 illustrates a plot of Limit of Detection (LOD) vs. Depth of Read (DOR) for detection of SNV Transition and Transversion mutations in a 84-plex PCR reaction at 25 PCR cycles.

[208] FIG. 62 is a plot illustrating detection of tumor mosaicism detected in single cells from the tumor.

[201] While the above-identified drawings set forth presently disclosed embodiments, other embodiments are also contemplated, as noted in the discussion. This disclosure presents illustrative embodiments by way of representation and not limitation. Numerous other modifications and embodiments can be devised by those skilled in the art which fall within the scope and spirit of the principles of the presently disclosed embodiments.

#### **DETAILED DESCRIPTION OF THE INVENTION**

[202] The present invention generally relates, at least in part, to improved methods of determining the presence or absence of copy number variations, such as deletions or duplications of chromosome segments or entire chromosomes. The methods are particularly useful for detecting small deletions or duplications, which can be difficult to detect with high specificity and sensitivity using prior methods due to the small amount of data available from the relevant chromosome segment. Methods of the invention can also be used to detect deletions or duplications that are only present in a small percentage of the cells or nucleic acid molecules that are tested. This allows deletions or duplications to be detected prior to the occurrence of disease (such as at a precancerous stage) or in the early stages of disease, such as before a large number of diseased cells (such as cancer cells) with the deletion or duplication accumulate. The more accurate detection of deletions or duplications associated with a disease or disorder enable improved methods for

diagnosing, prognosticating, preventing, delaying, stabilizing, or treating the disease or disorder. Several deletions or duplications are known to be associated with cancer or with severe mental or physical handicaps.

**[203]** Successful treatment of a disease such as cancer often relies on early diagnosis, correct staging of the disease, selection of an effective therapeutic regimen, and close monitoring to prevent or detect relapse. For cancer diagnosis, histological evaluation of tumor material obtained from tissue biopsy is often considered the most reliable method. However, the invasive nature of biopsy-based sampling has rendered it impractical for mass screening and regular follow up. Therefore, the present methods have the advantage of being able to be performed non-invasively if desired for relatively low cost with fast turnaround time. The targeted sequencing that may be used by the methods of the invention requires less reads than shotgun sequencing, such as a few million reads instead of 40 million reads, thereby decreasing cost. The multiplex PCR and next generation sequencing that may be used increase throughput and reduces costs.

**[204]** In some embodiments, the methods are used to detect a deletion or duplication in an individual. A sample from the individual that contains cells or nucleic acids suspected of having a deletion or duplication may be analyzed. In some embodiments, the sample is from a tissue or organ suspected of having a deletion or duplication, such as cells or a mass suspected of being cancerous. The methods of the invention can be used to detect deletions or duplications that are only present in one cell or a small number of cells in a mixture containing cells with the deletion or duplication and cells without the deletion or duplication. In some embodiments, cfDNA or cfRNA from a blood sample from the individual is analyzed. In some embodiments, cfDNA or cfRNA is secreted by cells, such as cancer cells. In some embodiments, cfDNA or cfRNA is released by cells undergoing necrosis or apoptosis, such as cancer cells. The methods of the invention can be used to detect deletions or duplications that are only present in a small percentage of the cfDNA or cfRNA. In some embodiments, one or more cells from an embryo are tested.

**[205]** In some embodiments, the methods are used for non-invasive or invasive prenatal testing of a fetus. These methods can be used to determine the presence or absence of deletions or duplications of a chromosome segment or an entire chromosome, such as deletions or duplications known to be associated severe mental or physical handicaps, learning disabilities, or cancer. In some embodiments for non-invasive prenatal testing (NIPT), cells, cfDNA or cfRNA from a blood sample from the pregnant mother is tested. The methods allow the detection of a deletion or duplication in the cells, cfDNA, or cfRNA from the fetus despite the large amount of cells, cfDNA, or cfRNA from the mother that is also present. In some

diagnosing, prognosticating, preventing, delaying, stabilizing, or treating the disease or disorder. Several deletions or duplications are known to be associated with cancer or with severe mental or physical handicaps.

**[203]** Successful treatment of a disease such as cancer often relies on early diagnosis, correct staging of the disease, selection of an effective therapeutic regimen, and close monitoring to prevent or detect relapse. For cancer diagnosis, histological evaluation of tumor material obtained from tissue biopsy is often considered the most reliable method. However, the invasive nature of biopsy-based sampling has rendered it impractical for mass screening and regular follow up. Therefore, the present methods have the advantage of being able to be performed non-invasively if desired for relatively low cost with fast turnaround time. The targeted sequencing that may be used by the methods of the invention requires less reads than shotgun sequencing, such as a few million reads instead of 40 million reads, thereby decreasing cost. The multiplex PCR and next generation sequencing that may be used increase throughput and reduces costs.

**[204]** In some embodiments, the methods are used to detect a deletion or duplication in an individual. A sample from the individual that contains cells or nucleic acids suspected of having a deletion or duplication may be analyzed. In some embodiments, the sample is from a tissue or organ suspected of having a deletion or duplication, such as cells or a mass suspected of being cancerous. The methods of the invention can be used to detect deletions or duplications that are only present in one cell or a small number of cells in a mixture containing cells with the deletion or duplication and cells without the deletion or duplication. In some embodiments, cfDNA or cfRNA from a blood sample from the individual is analyzed. In some embodiments, cfDNA or cfRNA is secreted by cells, such as cancer cells. In some embodiments, cfDNA or cfRNA is released by cells undergoing necrosis or apoptosis, such as cancer cells. The methods of the invention can be used to detect deletions or duplications that are only present in a small percentage of the cfDNA or cfRNA. In some embodiments, one or more cells from an embryo are tested.

**[205]** In some embodiments, the methods are used for non-invasive or invasive prenatal testing of a fetus. These methods can be used to determine the presence or absence of deletions or duplications of a chromosome segment or an entire chromosome, such as deletions or duplications known to be associated severe mental or physical handicaps, learning disabilities, or cancer. In some embodiments for non-invasive prenatal testing (NIPT), cells, cfDNA or cfRNA from a blood sample from the pregnant mother is tested. The methods allow the detection of a deletion or duplication in the cells, cfDNA, or cfRNA from the fetus despite the large amount of cells, cfDNA, or cfRNA from the mother that is also present. In some

embodiments for invasive prenatal testing, DNA or RNA from a sample from the fetus is tested (such as a CVS or amniocentesis sample). Even if the sample is contaminated with DNA or RNA from the pregnant mother, the methods can be used to detect a deletion or duplication in the fetal DNA or RNA.

**[206]** In addition to determining the presence or absence of copy number variation, one or more other factors can be analyzed if desired. These factors can be used to increase the accuracy of the diagnosis (such as determining the presence or absence of cancer or an increased risk for cancer, classifying the cancer, or staging the cancer) or prognosis. These factors can also be used to select a particular therapy or treatment regimen that is likely to be effective in the subject. Exemplary factors include the presence or absence of polymorphisms or mutation; altered (increased or decreased) levels of total or particular cfDNA, cfRNA, microRNA (miRNA); altered (increased or decreased) tumor fraction; altered (increased or decreased) methylation levels, altered (increased or decreased) DNA integrity, altered (increased or decreased) or alternative mRNA splicing.

**[207]** The following sections describe methods for detecting deletions or duplications using phased data (such as inferred or measured phased data) or unphased data; samples that can be tested; methods for sample preparation, amplification, and quantification; methods for phasing genetic data; polymorphisms, mutations, nucleic acid alterations, mRNA splicing alterations, and changes in nucleic acid levels that can be detected; databases with results from the methods, other risk factors and screening methods; cancers that can be diagnosed or treated; cancer treatments; cancer models for testing treatments; and methods for formulating and administering treatments.

#### Exemplary Methods for Detecting Deletions and Duplications Using Phased Data

**[208]** Some of the methods of the invention are based in part on the discovery that using phased data for detecting CNVs decreases the false negative and false positive rates compared to using unphased data (Figures 20A-27). This improvement is greatest for samples with CNVs present in low levels. Thus, phase data increases the accuracy of CNV detection compared to using unphased data (such as methods that calculate allele ratios at one or more loci or aggregate allele ratios to give an aggregated value (such as an average value) over a chromosome or chromosome segment without considering whether the allele ratios at different loci indicate that the same or different haplotypes appear to be present in an abnormal amount). Using phased data allows a more accurate determination to be made of whether differences between measured and expected allele ratios are due to noise or due to the presence of a CNV. For example, if the differences between measured and expected allele ratios at most or all of the loci in a region indicate that

embodiments for invasive prenatal testing, DNA or RNA from a sample from the fetus is tested (such as a CVS or amniocentesis sample). Even if the sample is contaminated with DNA or RNA from the pregnant mother, the methods can be used to detect a deletion or duplication in the fetal DNA or RNA.

**[206]** In addition to determining the presence or absence of copy number variation, one or more other factors can be analyzed if desired. These factors can be used to increase the accuracy of the diagnosis (such as determining the presence or absence of cancer or an increased risk for cancer, classifying the cancer, or staging the cancer) or prognosis. These factors can also be used to select a particular therapy or treatment regimen that is likely to be effective in the subject. Exemplary factors include the presence or absence of polymorphisms or mutation; altered (increased or decreased) levels of total or particular cfDNA, cfRNA, microRNA (miRNA); altered (increased or decreased) tumor fraction; altered (increased or decreased) methylation levels, altered (increased or decreased) DNA integrity, altered (increased or decreased) or alternative mRNA splicing.

**[207]** The following sections describe methods for detecting deletions or duplications using phased data (such as inferred or measured phased data) or unphased data; samples that can be tested; methods for sample preparation, amplification, and quantification; methods for phasing genetic data; polymorphisms, mutations, nucleic acid alterations, mRNA splicing alterations, and changes in nucleic acid levels that can be detected; databases with results from the methods, other risk factors and screening methods; cancers that can be diagnosed or treated; cancer treatments; cancer models for testing treatments; and methods for formulating and administering treatments.

#### Exemplary Methods for Detecting Deletions and Duplications Using Phased Data

**[208]** Some of the methods of the invention are based in part on the discovery that using phased data for detecting CNVs decreases the false negative and false positive rates compared to using unphased data (Figures 20A-27). This improvement is greatest for samples with CNVs present in low levels. Thus, phase data increases the accuracy of CNV detection compared to using unphased data (such as methods that calculate allele ratios at one or more loci or aggregate allele ratios to give an aggregated value (such as an average value) over a chromosome or chromosome segment without considering whether the allele ratios at different loci indicate that the same or different haplotypes appear to be present in an abnormal amount). Using phased data allows a more accurate determination to be made of whether differences between measured and expected allele ratios are due to noise or due to the presence of a CNV. For example, if the differences between measured and expected allele ratios at most or all of the loci in a region indicate that

the same haplotype is overrepresented, then a CNV is more likely to be present. Using linkage between alleles in a haplotype allows one to determine whether the measured genetic data is consistent with the same haplotype being overrepresented (rather than random noise). In contrast, if the differences between measured and expected allele ratios are only due to noise (such as experimental error), then in some embodiments, about half the time the first haplotype appears to be overrepresented and about the other half of the time, the second haplotype appears to be overrepresented.

**[209]** Accuracy can be increased by taking into account the linkage between SNPs, and the likelihood of crossovers having occurred during the meiosis that gave rise to the gametes that formed the embryo that grew into the fetus. Using linkage when creating the expected distribution of allele measurements for one or more hypotheses allows the creation of expected allele measurements distributions that correspond to reality considerably better than when linkage is not used. For example, imagine that there are two SNPs, 1 and 2 located nearby one another, and the mother is A at SNP 1 and A at SNP 2 on one homolog, and B at SNP 1 and B at SNP 2 on homo log two. If the father is A for both SNPs on both homo logs, and a B is measured for the fetus SNP 1, this indicates that homo log two has been inherited by the fetus, and therefore that there is a much higher likelihood of a B being present in the fetus at SNP 2. A model that takes into account linkage can predict this, while a model that does not take linkage into account cannot. Alternately, if a mother is AB at SNP 1 and AB at nearby SNP 2, then two hypotheses corresponding to maternal trisomy at that location can be used - one involving a matching copy error (nondisjunction in meiosis II or mitosis in early fetal development), and one involving an unmatching copy error (nondisjunction in meiosis I). In the case of a matching copy error trisomy, if the fetus inherited an AA from the mother at SNP 1, then the fetus is much more likely to inherit either an AA or BB from the mother at SNP 2, but not AB. In the case of an unmatching copy error, the fetus inherits an AB from the mother at both SNPs. The allele distribution hypotheses made by a CNV calling method that takes into account linkage can makes these predictions, and therefore correspond to the actual allele measurements to a considerably greater extent than a CNV calling method that does not take into account linkage.

**[210]** In some embodiments, phased genetic data is used to determine if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of an individual (such as in the genome of one or more cells or in cfDNA or cfRNA). Exemplary overrepresentations include the duplication of the first homologous chromosome segment or the deletion of the second homologous chromosome segment. In some embodiments, there is not an overrepresentation since the first and homologous chromosome segments are

the same haplotype is overrepresented, then a CNV is more likely to be present. Using linkage between alleles in a haplotype allows one to determine whether the measured genetic data is consistent with the same haplotype being overrepresented (rather than random noise). In contrast, if the differences between measured and expected allele ratios are only due to noise (such as experimental error), then in some embodiments, about half the time the first haplotype appears to be overrepresented and about the other half of the time, the second haplotype appears to be overrepresented.

**[209]** Accuracy can be increased by taking into account the linkage between SNPs, and the likelihood of crossovers having occurred during the meiosis that gave rise to the gametes that formed the embryo that grew into the fetus. Using linkage when creating the expected distribution of allele measurements for one or more hypotheses allows the creation of expected allele measurements distributions that correspond to reality considerably better than when linkage is not used. For example, imagine that there are two SNPs, 1 and 2 located nearby one another, and the mother is A at SNP 1 and A at SNP 2 on one homolog, and B at SNP 1 and B at SNP 2 on homolog two. If the father is A for both SNPs on both homologs, and a B is measured for the fetus SNP 1, this indicates that homolog two has been inherited by the fetus, and therefore that there is a much higher likelihood of a B being present in the fetus at SNP 2. A model that takes into account linkage can predict this, while a model that does not take linkage into account cannot. Alternately, if a mother is AB at SNP 1 and AB at nearby SNP 2, then two hypotheses corresponding to maternal trisomy at that location can be used - one involving a matching copy error (nondisjunction in meiosis II or mitosis in early fetal development), and one involving an unmatching copy error (nondisjunction in meiosis I). In the case of a matching copy error trisomy, if the fetus inherited an AA from the mother at SNP 1, then the fetus is much more likely to inherit either an AA or BB from the mother at SNP 2, but not AB. In the case of an unmatching copy error, the fetus inherits an AB from the mother at both SNPs. The allele distribution hypotheses made by a CNV calling method that takes into account linkage can make these predictions, and therefore correspond to the actual allele measurements to a considerably greater extent than a CNV calling method that does not take into account linkage.

**[210]** In some embodiments, phased genetic data is used to determine if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of an individual (such as in the genome of one or more cells or in cfDNA or cfRNA). Exemplary overrepresentations include the duplication of the first homologous chromosome segment or the deletion of the second homologous chromosome segment. In some embodiments, there is not an overrepresentation since the first and homologous chromosome segments are

present in equal proportions (such as one copy of each segment in a diploid sample). In some embodiments, calculated allele ratios in a nucleic acid sample are compared to expected allele ratios to determine if there is an overrepresentation as described further below.

**[211]** In some embodiments, the method includes obtaining phased genetic data for the first homologous chromosome segment comprising the identity of the allele present at that locus on the first homologous chromosome segment for each locus in a set of polymorphic loci on the first homologous chromosome segment, obtaining phased genetic data for the second homologous chromosome segment comprising the identity of the allele present at that locus on the second homologous chromosome segment for each locus in the set of polymorphic loci on the second homologous chromosome segment, and obtaining measured genetic allelic data comprising, for each of the alleles at each of the loci in the set of polymorphic loci, the amount of each allele present in a sample of DNA or RNA from one or more target cells and one or more non-target cells from the individual. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment; calculating, for each of the hypotheses, expected genetic data for the plurality of loci in the sample from the obtained phased genetic data for one or more possible ratios of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample; calculating (such as calculating on a computer) for each possible ratio of DNA or RNA and for each hypothesis, the data fit between the obtained genetic data of the sample and the expected genetic data for the sample for that possible ratio of DNA or RNA and for that hypothesis; ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more cells from the individual.

**[212]** In one aspect, the invention features a method for determining a number of copies of a chromosome or chromosome segment of interest in the genome of a fetus. In some embodiments, the method includes obtaining phased genetic data for at least one biological parent of the fetus, wherein the phased genetic data comprises the identity of the allele present for each locus in a set of polymorphic loci on a first homologous chromosome segment and a second homologous chromosome segment in the parent. In some embodiments, the method includes obtaining genetic data at the set of polymorphic loci on the chromosome or chromosome segment in a mixed sample of DNA or RNA comprising fetal DNA or RNA and maternal DNA or RNA from the mother of the fetus by measuring the quantity of each allele at each locus. In some embodiments, the method includes enumerating a set of one

present in equal proportions (such as one copy of each segment in a diploid sample). In some embodiments, calculated allele ratios in a nucleic acid sample are compared to expected allele ratios to determine if there is an overrepresentation as described further below.

**[211]** In some embodiments, the method includes obtaining phased genetic data for the first homologous chromosome segment comprising the identity of the allele present at that locus on the first homologous chromosome segment for each locus in a set of polymorphic loci on the first homologous chromosome segment, obtaining phased genetic data for the second homologous chromosome segment comprising the identity of the allele present at that locus on the second homologous chromosome segment for each locus in the set of polymorphic loci on the second homologous chromosome segment, and obtaining measured genetic allelic data comprising, for each of the alleles at each of the loci in the set of polymorphic loci, the amount of each allele present in a sample of DNA or RNA from one or more target cells and one or more non-target cells from the individual. In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment; calculating, for each of the hypotheses, expected genetic data for the plurality of loci in the sample from the obtained phased genetic data for one or more possible ratios of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample; calculating (such as calculating on a computer) for each possible ratio of DNA or RNA and for each hypothesis, the data fit between the obtained genetic data of the sample and the expected genetic data for the sample for that possible ratio of DNA or RNA and for that hypothesis; ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more cells from the individual.

**[212]** In one aspect, the invention features a method for determining a number of copies of a chromosome or chromosome segment of interest in the genome of a fetus. In some embodiments, the method includes obtaining phased genetic data for at least one biological parent of the fetus, wherein the phased genetic data comprises the identity of the allele present for each locus in a set of polymorphic loci on a first homologous chromosome segment and a second homologous chromosome segment in the parent. In some embodiments, the method includes obtaining genetic data at the set of polymorphic loci on the chromosome or chromosome segment in a mixed sample of DNA or RNA comprising fetal DNA or RNA and maternal DNA or RNA from the mother of the fetus by measuring the quantity of each allele at each locus. In some embodiments, the method includes enumerating a set of one

or more hypotheses specifying the number of copies of the chromosome or chromosome segment of interest present in the genome of the fetus. In some embodiments, the method includes creating (such as creating on a computer) for each of the hypotheses, a probability distribution of the expected quantity of each allele at each of the plurality of loci in mixed sample from the (i) the obtained phased genetic data from the parent(s) and optionally (ii) the probability of one or more crossovers that may have occurred during the formation of a gamete that contributed a copy of the chromosome or chromosome segment of interest to the fetus; calculating (such as calculating on a computer) a fit, for each of the hypotheses, between (1) the obtained genetic data of the mixed sample and (2) the probability distribution of the expected quantity of each allele at each of the plurality of loci in mixed sample for that hypothesis; ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the number of copies of the chromosome segment of interest in the genome of the fetus.

**[213]** In some embodiments, the method involves obtaining phased genetic data using any of the methods described herein or any known method. In some embodiments, the method involves simultaneously or sequentially in any order (i) obtaining phased genetic data for the first homologous chromosome segment comprising the identity of the allele present at that locus on the first homologous chromosome segment for each locus in a set of polymorphic loci on the first homologous chromosome segment, (ii) obtaining phased genetic data for the second homologous chromosome segment comprising the identity of the allele present at that locus on the second homologous chromosome segment for each locus in the set of polymorphic loci on the second homologous chromosome segment, and (iii) obtaining measured genetic allelic data comprising the amount of each allele at each of the loci in the set of polymorphic loci in a sample of DNA from one or more cells from the individual.

**[214]** In some embodiments, the method involves calculating allele ratios for one or more loci in the set of polymorphic loci that are heterozygous in at least one cell from which the sample was derived (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother). In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus. In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles (such as the allele on the first homologous chromosome segment) divided by the measured quantity of one or more other alleles (such as the allele on the second homologous chromosome segment) for the locus. The calculated allele ratios may be calculated using any of the methods described herein or any standard method (such as any

or more hypotheses specifying the number of copies of the chromosome or chromosome segment of interest present in the genome of the fetus. In some embodiments, the method includes creating (such as creating on a computer) for each of the hypotheses, a probability distribution of the expected quantity of each allele at each of the plurality of loci in mixed sample from the (i) the obtained phased genetic data from the parent(s) and optionally (ii) the probability of one or more crossovers that may have occurred during the formation of a gamete that contributed a copy of the chromosome or chromosome segment of interest to the fetus; calculating (such as calculating on a computer) a fit, for each of the hypotheses, between (1) the obtained genetic data of the mixed sample and (2) the probability distribution of the expected quantity of each allele at each of the plurality of loci in mixed sample for that hypothesis; ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the number of copies of the chromosome segment of interest in the genome of the fetus.

**[213]** In some embodiments, the method involves obtaining phased genetic data using any of the methods described herein or any known method. In some embodiments, the method involves simultaneously or sequentially in any order (i) obtaining phased genetic data for the first homologous chromosome segment comprising the identity of the allele present at that locus on the first homologous chromosome segment for each locus in a set of polymorphic loci on the first homologous chromosome segment, (ii) obtaining phased genetic data for the second homologous chromosome segment comprising the identity of the allele present at that locus on the second homologous chromosome segment for each locus in the set of polymorphic loci on the second homologous chromosome segment, and (iii) obtaining measured genetic allelic data comprising the amount of each allele at each of the loci in the set of polymorphic loci in a sample of DNA from one or more cells from the individual.

**[214]** In some embodiments, the method involves calculating allele ratios for one or more loci in the set of polymorphic loci that are heterozygous in at least one cell from which the sample was derived (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother). In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus. In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles (such as the allele on the first homologous chromosome segment) divided by the measured quantity of one or more other alleles (such as the allele on the second homologous chromosome segment) for the locus. The calculated allele ratios may be calculated using any of the methods described herein or any standard method (such as any

mathematical transformation of the calculated allele ratios described herein).

**[215]** In some embodiments, the method involves determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment by comparing one or more calculated allele ratios for a locus to an allele ratio that is expected for that locus if the first and second homologous chromosome segments are present in equal proportions. In some embodiments, the expected allele ratio assumes the possible alleles for a locus have an equal likelihood of being present. In some embodiments in which the calculated allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus, the corresponding expected allele ratio is 0.5 for a biallelic locus, or 1/3 for a triallelic locus. In some embodiments, the expected allele ratio is the same for all the loci, such as 0.5 for all loci. In some embodiments, the expected allele ratio assumes that the possible alleles for a locus can have a different likelihood of being present, such as the likelihood based on the frequency of each of the alleles in a particular population that the subject belongs in, such as a population based on the ancestry of the subject. Such allele frequencies are publicly available (*see, e.g.*, HapMap Project; Perlegen Human Haplotype Project; web at [ncbi.nlm.nih.gov/projects/SNP/](http://ncbi.nlm.nih.gov/projects/SNP/); Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001 Jan 1 ;29(1):308-11, which are each incorporated by reference in its entirety). In some embodiments, the expected allele ratio is the allele ratio that is expected for the particular individual being tested for a particular hypothesis specifying the degree of overrepresentation of the first homologous chromosome segment. For example, the expected allele ratio for a particular individual may be determined based on phased or unphased genetic data from the individual (such as from a sample from the individual that is unlikely to have a deletion or duplication such as a noncancerous sample) or data from one or more relatives from the individual. In some embodiments for prenatal testing, the expected allele ratio is the allele ratio that is expected for a mixed sample that includes DNA or RNA from the pregnant mother and the fetus (such as a maternal plasma or serum sample that includes cfDNA from the mother and cfDNA from the fetus) for a particular hypothesis specifying the degree of overrepresentation of the first homologous chromosome segment. For example, the expected allele ratio for the mixed sample may be determined based on genetic data from the mother and predicted genetic data for the fetus (such as predictions for alleles that the fetus may have inherited from the mother and/or father). In some embodiments, phased or unphased genetic data from a sample of DNA or RNA from only the mother (such as the buffy coat from a maternal blood sample) is to determine the alleles from the maternal DNA or RNA in the mixed sample as well as alleles that the fetus may have been inherited from the mother (and

mathematical transformation of the calculated allele ratios described herein).

**[215]** In some embodiments, the method involves determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment by comparing one or more calculated allele ratios for a locus to an allele ratio that is expected for that locus if the first and second homologous chromosome segments are present in equal proportions. In some embodiments, the expected allele ratio assumes the possible alleles for a locus have an equal likelihood of being present. In some embodiments in which the calculated allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus, the corresponding expected allele ratio is 0.5 for a biallelic locus, or 1/3 for a triallelic locus. In some embodiments, the expected allele ratio is the same for all the loci, such as 0.5 for all loci. In some embodiments, the expected allele ratio assumes that the possible alleles for a locus can have a different likelihood of being present, such as the likelihood based on the frequency of each of the alleles in a particular population that the subject belongs in, such as a population based on the ancestry of the subject. Such allele frequencies are publicly available (*see, e.g.*, HapMap Project; Perlegen Human Haplotype Project; web at [ncbi.nlm.nih.gov/projects/SNP/](http://ncbi.nlm.nih.gov/projects/SNP/); Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001 Jan 1 ;29(1):308-11, which are each incorporated by reference in its entirety). In some embodiments, the expected allele ratio is the allele ratio that is expected for the particular individual being tested for a particular hypothesis specifying the degree of overrepresentation of the first homologous chromosome segment. For example, the expected allele ratio for a particular individual may be determined based on phased or unphased genetic data from the individual (such as from a sample from the individual that is unlikely to have a deletion or duplication such as a noncancerous sample) or data from one or more relatives from the individual. In some embodiments for prenatal testing, the expected allele ratio is the allele ratio that is expected for a mixed sample that includes DNA or RNA from the pregnant mother and the fetus (such as a maternal plasma or serum sample that includes cfDNA from the mother and cfDNA from the fetus) for a particular hypothesis specifying the degree of overrepresentation of the first homologous chromosome segment. For example, the expected allele ratio for the mixed sample may be determined based on genetic data from the mother and predicted genetic data for the fetus (such as predictions for alleles that the fetus may have inherited from the mother and/or father). In some embodiments, phased or unphased genetic data from a sample of DNA or RNA from only the mother (such as the buffy coat from a maternal blood sample) is to determine the alleles from the maternal DNA or RNA in the mixed sample as well as alleles that the fetus may have been inherited from the mother (and

thus may be present in the fetal DNA or RNA in the mixed sample). In some embodiments, phased or unphased genetic data from a sample of DNA or RNA from only the father is used to determine the alleles that the fetus may have been inherited from the father (and thus may be present in the fetal DNA or RNA in the mixed sample). The expected allele ratios may be calculated using any of the methods described herein or any standard method (such as any mathematical transformation of the expected allele ratios described herein) (U.S. Publication No 2012/0270212, filed Nov. 18, 2011, which is hereby incorporated by reference in its entirety).

**[216]** In some embodiments, a calculated allele ratio is indicative of an overrepresentation of the number of copies of the first homologous chromosome segment if either (i) the allele ratio for the measured quantity of the allele present at that locus on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is greater than the expected allele ratio for that locus, or (ii) the allele ratio for the measured quantity of the allele present at that locus on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than the expected allele ratio for that locus. In some embodiments, a calculated allele ratio is only considered indicative of overrepresentation if it is significantly greater or lower than the expected ratio for that locus. In some embodiments, a calculated allele ratio is indicative of no overrepresentation of the number of copies of the first homologous chromosome segment if either (i) the allele ratio for the measured quantity of the allele present at that locus on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than or equal to the expected allele ratio for that locus, or (ii) the allele ratio for the measured quantity of the allele present at that locus on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is greater than or equal to the expected allele ratio for that locus. In some embodiments, calculated ratios equal to the corresponding expected ratio are ignored (since they are indicative of no overrepresentation).

**[217]** In various embodiments, one or more of the following methods is used to compare one or more of the calculated allele ratios to the corresponding expected allele ratio(s). In some embodiments, one determines whether the calculated allele ratio is above or below the expected allele ratio for a particular locus irrespective of the magnitude of the difference. In some embodiments, one determines the magnitude of the difference between the calculated allele ratio and the expected allele ratio for a particular locus irrespective of whether the calculated allele ratio is above or below the expected allele ratio. In some embodiments, one determines whether the calculated allele ratio is above or below the expected allele ratio and the magnitude of the difference for a particular locus. In some embodiments, one determines

thus may be present in the fetal DNA or RNA in the mixed sample). In some embodiments, phased or unphased genetic data from a sample of DNA or RNA from only the father is used to determine the alleles that the fetus may have been inherited from the father (and thus may be present in the fetal DNA or RNA in the mixed sample). The expected allele ratios may be calculated using any of the methods described herein or any standard method (such as any mathematical transformation of the expected allele ratios described herein) (U.S. Publication No 2012/0270212, filed Nov. 18, 2011, which is hereby incorporated by reference in its entirety).

**[216]** In some embodiments, a calculated allele ratio is indicative of an overrepresentation of the number of copies of the first homologous chromosome segment if either (i) the allele ratio for the measured quantity of the allele present at that locus on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is greater than the expected allele ratio for that locus, or (ii) the allele ratio for the measured quantity of the allele present at that locus on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than the expected allele ratio for that locus. In some embodiments, a calculated allele ratio is only considered indicative of overrepresentation if it is significantly greater or lower than the expected ratio for that locus. In some embodiments, a calculated allele ratio is indicative of no overrepresentation of the number of copies of the first homologous chromosome segment if either (i) the allele ratio for the measured quantity of the allele present at that locus on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than or equal to the expected allele ratio for that locus, or (ii) the allele ratio for the measured quantity of the allele present at that locus on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is greater than or equal to the expected allele ratio for that locus. In some embodiments, calculated ratios equal to the corresponding expected ratio are ignored (since they are indicative of no overrepresentation).

**[217]** In various embodiments, one or more of the following methods is used to compare one or more of the calculated allele ratios to the corresponding expected allele ratio(s). In some embodiments, one determines whether the calculated allele ratio is above or below the expected allele ratio for a particular locus irrespective of the magnitude of the difference. In some embodiments, one determines the magnitude of the difference between the calculated allele ratio and the expected allele ratio for a particular locus irrespective of whether the calculated allele ratio is above or below the expected allele ratio. In some embodiments, one determines whether the calculated allele ratio is above or below the expected allele ratio and the magnitude of the difference for a particular locus. In some embodiments, one determines

whether the average or weighted average value of the calculated allele ratios is above or below the average or weighted average value of the expected allele ratios irrespective of the magnitude of the difference. In some embodiments, one determines the magnitude of the difference between the average or weighted average value of the calculated allele ratios and the average or weighted average value of the expected allele ratios irrespective of whether the average or weighted average of the calculated allele ratio is above or below the average or weighted average value of the expected allele ratio. In some embodiments, one determines whether the average or weighted average value of the calculated allele ratios is above or below the average or weighted average value of the expected allele ratios and the magnitude of the difference. In some embodiments, one determines an average or weighted average value of the magnitude of the difference between the calculated allele ratios and the expected allele ratios.

**[218]** In some embodiments, the magnitude of the difference between the calculated allele ratio and the expected allele ratio for one or more loci is used to determine whether the overrepresentation of the number of copies of the first homologous chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment in the genome of one or more of the cells.

**[219]** In some embodiments, an overrepresentation of the number of copies of the first homologous chromosome segment is determined to be present if one or more of following conditions is met. In some embodiments, the number of calculated allele ratios that are indicative of an overrepresentation of the number of copies of the first homologous chromosome segment is above a threshold value. In some embodiments, the number of calculated allele ratios that are indicative of no overrepresentation of the number of copies of the first homologous chromosome segment is below a threshold value. In some embodiments, the magnitude of the difference between the calculated allele ratios that are indicative of an overrepresentation of the number of copies of the first homologous chromosome segment and the corresponding expected allele ratios is above a threshold value. In some embodiments, for all calculated allele ratios that are indicative of overrepresentation, the sum of the magnitude of the difference between a calculated allele ratio and the corresponding expected allele ratio is above a threshold value. In some embodiments, the magnitude of the difference between the calculated allele ratios that are indicative of no overrepresentation of the number of copies of the first homologous chromosome segment and the corresponding expected allele ratios is below a threshold value. In some embodiments, the average or weighted average value of the calculated allele ratios for the measured quantity of the allele present on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is

whether the average or weighted average value of the calculated allele ratios is above or below the average or weighted average value of the expected allele ratios irrespective of the magnitude of the difference. In some embodiments, one determines the magnitude of the difference between the average or weighted average value of the calculated allele ratios and the average or weighted average value of the expected allele ratios irrespective of whether the average or weighted average of the calculated allele ratio is above or below the average or weighted average value of the expected allele ratio. In some embodiments, one determines whether the average or weighted average value of the calculated allele ratios is above or below the average or weighted average value of the expected allele ratios and the magnitude of the difference. In some embodiments, one determines an average or weighted average value of the magnitude of the difference between the calculated allele ratios and the expected allele ratios.

**[218]** In some embodiments, the magnitude of the difference between the calculated allele ratio and the expected allele ratio for one or more loci is used to determine whether the overrepresentation of the number of copies of the first homologous chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment in the genome of one or more of the cells.

**[219]** In some embodiments, an overrepresentation of the number of copies of the first homologous chromosome segment is determined to be present if one or more of following conditions is met. In some embodiments, the number of calculated allele ratios that are indicative of an overrepresentation of the number of copies of the first homologous chromosome segment is above a threshold value. In some embodiments, the number of calculated allele ratios that are indicative of no overrepresentation of the number of copies of the first homologous chromosome segment is below a threshold value. In some embodiments, the magnitude of the difference between the calculated allele ratios that are indicative of an overrepresentation of the number of copies of the first homologous chromosome segment and the corresponding expected allele ratios is above a threshold value. In some embodiments, for all calculated allele ratios that are indicative of overrepresentation, the sum of the magnitude of the difference between a calculated allele ratio and the corresponding expected allele ratio is above a threshold value. In some embodiments, the magnitude of the difference between the calculated allele ratios that are indicative of no overrepresentation of the number of copies of the first homologous chromosome segment and the corresponding expected allele ratios is below a threshold value. In some embodiments, the average or weighted average value of the calculated allele ratios for the measured quantity of the allele present on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is

greater than the average or weighted average value of the expected allele ratios by at least a threshold value. In some embodiments, the average or weighted average value of the calculated allele ratios for the measured quantity of the allele present on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than the average or weighted average value of the expected allele ratios by at least a threshold value. In some embodiments, the data fit between the calculated allele ratios and allele ratios that are predicted for an overrepresentation of the number of copies of the first homologous chromosome segment is below a threshold value (indicative of a good data fit). In some embodiments, the data fit between the calculated allele ratios and allele ratios that are predicted for no overrepresentation of the number of copies of the first homologous chromosome segment is above a threshold value (indicative of a poor data fit).

[206] In some embodiments, an overrepresentation of the number of copies of the first homologous chromosome segment is determined to be absent if one or more of following conditions is met. In some embodiments, the number of calculated allele ratios that are indicative of an overrepresentation of the number of copies of the first homologous chromosome segment is below a threshold value. In some embodiments, the number of calculated allele ratios that are indicative of no overrepresentation of the number of copies of the first homologous chromosome segment is above a threshold value. In some embodiments, the magnitude of the difference between the calculated allele ratios that are indicative of an overrepresentation of the number of copies of the first homologous chromosome segment and the corresponding expected allele ratios is below a threshold value. In some embodiments, the magnitude of the difference between the calculated allele ratios that are indicative of no overrepresentation of the number of copies of the first homologous chromosome segment and the corresponding expected allele ratios is above a threshold value. In some embodiments, the average or weighted average value of the calculated allele ratios for the measured quantity of the allele present on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus minus the average or weighted average value of the expected allele ratios is less than a threshold value. In some embodiments, the average or weighted average value of the expected allele ratios minus the average or weighted average value of the calculated allele ratios for the measured quantity of the allele present on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than a threshold value. In some embodiments, the data fit between the calculated allele ratios and allele ratios that are predicted for an overrepresentation of the number of copies of the first homologous chromosome segment is above a threshold value. In some embodiments, the data fit between the calculated allele ratios and

greater than the average or weighted average value of the expected allele ratios by at least a threshold value. In some embodiments, the average or weighted average value of the calculated allele ratios for the measured quantity of the allele present on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than the average or weighted average value of the expected allele ratios by at least a threshold value. In some embodiments, the data fit between the calculated allele ratios and allele ratios that are predicted for an overrepresentation of the number of copies of the first homologous chromosome segment is below a threshold value (indicative of a good data fit). In some embodiments, the data fit between the calculated allele ratios and allele ratios that are predicted for no overrepresentation of the number of copies of the first homologous chromosome segment is above a threshold value (indicative of a poor data fit).

[206] In some embodiments, an overrepresentation of the number of copies of the first homologous chromosome segment is determined to be absent if one or more of following conditions is met. In some embodiments, the number of calculated allele ratios that are indicative of an overrepresentation of the number of copies of the first homologous chromosome segment is below a threshold value. In some embodiments, the number of calculated allele ratios that are indicative of no overrepresentation of the number of copies of the first homologous chromosome segment is above a threshold value. In some embodiments, the magnitude of the difference between the calculated allele ratios that are indicative of an overrepresentation of the number of copies of the first homologous chromosome segment and the corresponding expected allele ratios is below a threshold value. In some embodiments, the magnitude of the difference between the calculated allele ratios that are indicative of no overrepresentation of the number of copies of the first homologous chromosome segment and the corresponding expected allele ratios is above a threshold value. In some embodiments, the average or weighted average value of the calculated allele ratios for the measured quantity of the allele present on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus minus the average or weighted average value of the expected allele ratios is less than a threshold value. In some embodiments, the average or weighted average value of the expected allele ratios minus the average or weighted average value of the calculated allele ratios for the measured quantity of the allele present on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than a threshold value. In some embodiments, the data fit between the calculated allele ratios and allele ratios that are predicted for an overrepresentation of the number of copies of the first homologous chromosome segment is above a threshold value. In some embodiments, the data fit between the calculated allele ratios and

allele ratios that are predicted for no overrepresentation of the number of copies of the first homologous chromosome segment is below a threshold value. In some embodiments, the threshold is determined from empirical testing of samples known to have a CNV of interest and/or samples known to lack the CNV.

[207] In some embodiments, determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment. One exemplary hypothesis is the absence of an overrepresentation since the first and homologous chromosome segments are present in equal proportions (such as one copy of each segment in a diploid sample). Other exemplary hypotheses include the first homologous chromosome segment being duplicated one or more times (such as 1, 2, 3, 4, 5, or more extra copies of the first homologous chromosome compared to the number of copies of the second homologous chromosome segment). Another exemplary hypothesis includes the deletion of the second homologous chromosome segment. Yet another exemplary hypothesis is the deletion of both the first and the second homologous chromosome segments. In some embodiments, predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) are estimated for each hypothesis given the degree of overrepresentation specified by that hypothesis. In some embodiments, the likelihood that the hypothesis is correct is calculated by comparing the calculated allele ratios to the predicted allele ratios, and the hypothesis with the greatest likelihood is selected.

[208] In some embodiments, an expected distribution of a test statistic is calculated using the predicted allele ratios for each hypothesis. In some embodiments, the likelihood that the hypothesis is correct is calculated by comparing a test statistic that is calculated using the calculated allele ratios to the expected distribution of the test statistic that is calculated using the predicted allele ratios, and the hypothesis with the greatest likelihood is selected.

[209] In some embodiments, predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) are estimated given the phased genetic data for the first homologous chromosome segment, the phased genetic data for the second homologous chromosome segment, and the degree of overrepresentation specified by that hypothesis. In some embodiments, the likelihood that the hypothesis is correct is calculated by comparing the calculated allele ratios to the predicted allele ratios; and the hypothesis with the greatest likelihood is selected.

allele ratios that are predicted for no overrepresentation of the number of copies of the first homologous chromosome segment is below a threshold value. In some embodiments, the threshold is determined from empirical testing of samples known to have a CNV of interest and/or samples known to lack the CNV.

[207] In some embodiments, determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment. One exemplary hypothesis is the absence of an overrepresentation since the first and homologous chromosome segments are present in equal proportions (such as one copy of each segment in a diploid sample). Other exemplary hypotheses include the first homologous chromosome segment being duplicated one or more times (such as 1, 2, 3, 4, 5, or more extra copies of the first homologous chromosome compared to the number of copies of the second homologous chromosome segment). Another exemplary hypothesis includes the deletion of the second homologous chromosome segment. Yet another exemplary hypothesis is the deletion of both the first and the second homologous chromosome segments. In some embodiments, predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) are estimated for each hypothesis given the degree of overrepresentation specified by that hypothesis. In some embodiments, the likelihood that the hypothesis is correct is calculated by comparing the calculated allele ratios to the predicted allele ratios, and the hypothesis with the greatest likelihood is selected.

[208] In some embodiments, an expected distribution of a test statistic is calculated using the predicted allele ratios for each hypothesis. In some embodiments, the likelihood that the hypothesis is correct is calculated by comparing a test statistic that is calculated using the calculated allele ratios to the expected distribution of the test statistic that is calculated using the predicted allele ratios, and the hypothesis with the greatest likelihood is selected.

[209] In some embodiments, predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) are estimated given the phased genetic data for the first homologous chromosome segment, the phased genetic data for the second homologous chromosome segment, and the degree of overrepresentation specified by that hypothesis. In some embodiments, the likelihood that the hypothesis is correct is calculated by comparing the calculated allele ratios to the predicted allele ratios; and the hypothesis with the greatest likelihood is selected.

### *Use of Fraction of CNV or Disease in Mixed Samples*

[210] In some embodiments, the sample is a mixed sample with DNA or RNA from one or more target cells and one or more non-target cells. In some embodiments, the target cells are cells that have a CNV, such as a deletion or duplication of interest, and the non-target cells are cells that do not have the copy number variation of interest (such as a mixture of cells with the deletion or duplication of interest and cells without any of the deletions or duplications being tested). In some embodiments, the target cells are cells that are associated with a disease or disorder or an increased risk for disease or disorder (such as cancer cells), and the non-target cells are cells that are not associated with a disease or disorder or an increased risk for disease or disorder (such as noncancerous cells). In some embodiments, the target cells all have the same CNV. In some embodiments, two or more target cells have different CNVs. In some embodiments, one or more of the target cells has a CNV, polymorphism, or mutation associated with the disease or disorder or an increased risk for disease or disorder that is not found in at least one other target cell. In some such embodiments, the fraction of the cells that are associated with the disease or disorder or an increased risk for disease or disorder out of the total cells from a sample is assumed to be greater than or equal to the fraction of the most frequent of these CNVs, polymorphisms, or mutations in the sample. For example if 6% of the cells have a K-ras mutation, and 8% of the cells have a BRAF mutation, at least 8% of the cells are assumed to be cancerous.

[211] In some embodiments, the ratio of DNA (or RNA) from the one or more target cells to the total DNA (or RNA) in the sample is calculated. In some embodiments, a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment are enumerated. In some embodiments, predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) are estimated given the calculated ratio of DNA or RNA and the degree of overrepresentation specified by that hypothesis are estimated for each hypothesis. In some embodiments, the likelihood that the hypothesis is correct is calculated by comparing the calculated allele ratios to the predicted allele ratios, and the hypothesis with the greatest likelihood is selected.

[212] In some embodiments, an expected distribution of a test statistic calculated using the predicted allele ratios and the calculated ratio of DNA or RNA is estimated for each hypothesis. In some embodiments, the likelihood that the hypothesis is correct is determined by comparing a test statistic calculated using the calculated allele ratios and the calculated ratio of DNA or RNA to the expected distribution of the test statistic calculated using the predicted allele ratios and the calculated ratio of DNA

### *Use of Fraction of CNV or Disease in Mixed Samples*

[210] In some embodiments, the sample is a mixed sample with DNA or RNA from one or more target cells and one or more non-target cells. In some embodiments, the target cells are cells that have a CNV, such as a deletion or duplication of interest, and the non-target cells are cells that do not have the copy number variation of interest (such as a mixture of cells with the deletion or duplication of interest and cells without any of the deletions or duplications being tested). In some embodiments, the target cells are cells that are associated with a disease or disorder or an increased risk for disease or disorder (such as cancer cells), and the non-target cells are cells that are not associated with a disease or disorder or an increased risk for disease or disorder (such as noncancerous cells). In some embodiments, the target cells all have the same CNV. In some embodiments, two or more target cells have different CNVs. In some embodiments, one or more of the target cells has a CNV, polymorphism, or mutation associated with the disease or disorder or an increased risk for disease or disorder that is not found in at least one other target cell. In some such embodiments, the fraction of the cells that are associated with the disease or disorder or an increased risk for disease or disorder out of the total cells from a sample is assumed to be greater than or equal to the fraction of the most frequent of these CNVs, polymorphisms, or mutations in the sample. For example if 6% of the cells have a K-ras mutation, and 8% of the cells have a BRAF mutation, at least 8% of the cells are assumed to be cancerous.

[211] In some embodiments, the ratio of DNA (or RNA) from the one or more target cells to the total DNA (or RNA) in the sample is calculated. In some embodiments, a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment are enumerated. In some embodiments, predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) are estimated given the calculated ratio of DNA or RNA and the degree of overrepresentation specified by that hypothesis are estimated for each hypothesis. In some embodiments, the likelihood that the hypothesis is correct is calculated by comparing the calculated allele ratios to the predicted allele ratios, and the hypothesis with the greatest likelihood is selected.

[212] In some embodiments, an expected distribution of a test statistic calculated using the predicted allele ratios and the calculated ratio of DNA or RNA is estimated for each hypothesis. In some embodiments, the likelihood that the hypothesis is correct is determined by comparing a test statistic calculated using the calculated allele ratios and the calculated ratio of DNA or RNA to the expected distribution of the test statistic calculated using the predicted allele ratios and the calculated ratio of DNA

or RNA, and the hypothesis with the greatest likelihood is selected.

[213] In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment. In some embodiments, the method includes estimating, for each hypothesis, either (i) predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) given the degree of overrepresentation specified by that hypothesis or (ii) for one or more possible ratios of DNA or RNA, an expected distribution of a test statistic calculated using the predicted allele ratios and the possible ratio of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample. In some embodiments, a data fit is calculated by comparing either (i) the calculated allele ratios to the predicted allele ratios, or (ii) a test statistic calculated using the calculated allele ratios and the possible ratio of DNA or RNA to the expected distribution of the test statistic calculated using the predicted allele ratios and the possible ratio of DNA or RNA. In some embodiments, one or more of the hypotheses are ranked according to the data fit, and the hypothesis that is ranked the highest is selected. In some embodiments, a technique or algorithm, such as a search algorithm, is used for one or more of the following steps: calculating the data fit, ranking the hypotheses, or selecting the hypothesis that is ranked the highest. In some embodiments, the data fit is a fit to a beta-binomial distribution or a fit to a binomial distribution. In some embodiments, the technique or algorithm is selected from the group consisting of maximum likelihood estimation, maximum a-posteriori estimation, Bayesian estimation, dynamic estimation (such as dynamic Bayesian estimation), and expectation- maximization estimation. In some embodiments, the method includes applying the technique or algorithm to the obtained genetic data and the expected genetic data.

[214] In some embodiments, the method includes creating a partition of possible ratios that range from a lower limit to an upper limit for the ratio of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample. In some embodiments, a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment are enumerated. In some embodiments, the method includes estimating, for each of the possible ratios of DNA or RNA in the partition and for each hypothesis, either (i) predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) given the possible ratio of DNA or RNA and the degree of overrepresentation specified by that hypothesis or (ii) an expected distribution of a test statistic calculated using the predicted allele ratios and the possible ratio of DNA or RNA. In some embodiments, the method includes calculating, for

or RNA, and the hypothesis with the greatest likelihood is selected.

[213] In some embodiments, the method includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment. In some embodiments, the method includes estimating, for each hypothesis, either (i) predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) given the degree of overrepresentation specified by that hypothesis or (ii) for one or more possible ratios of DNA or RNA, an expected distribution of a test statistic calculated using the predicted allele ratios and the possible ratio of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample. In some embodiments, a data fit is calculated by comparing either (i) the calculated allele ratios to the predicted allele ratios, or (ii) a test statistic calculated using the calculated allele ratios and the possible ratio of DNA or RNA to the expected distribution of the test statistic calculated using the predicted allele ratios and the possible ratio of DNA or RNA. In some embodiments, one or more of the hypotheses are ranked according to the data fit, and the hypothesis that is ranked the highest is selected. In some embodiments, a technique or algorithm, such as a search algorithm, is used for one or more of the following steps: calculating the data fit, ranking the hypotheses, or selecting the hypothesis that is ranked the highest. In some embodiments, the data fit is a fit to a beta-binomial distribution or a fit to a binomial distribution. In some embodiments, the technique or algorithm is selected from the group consisting of maximum likelihood estimation, maximum a-posteriori estimation, Bayesian estimation, dynamic estimation (such as dynamic Bayesian estimation), and expectation- maximization estimation. In some embodiments, the method includes applying the technique or algorithm to the obtained genetic data and the expected genetic data.

[214] In some embodiments, the method includes creating a partition of possible ratios that range from a lower limit to an upper limit for the ratio of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample. In some embodiments, a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment are enumerated. In some embodiments, the method includes estimating, for each of the possible ratios of DNA or RNA in the partition and for each hypothesis, either (i) predicted allele ratios for the loci that are heterozygous in at least one cell (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother) given the possible ratio of DNA or RNA and the degree of overrepresentation specified by that hypothesis or (ii) an expected distribution of a test statistic calculated using the predicted allele ratios and the possible ratio of DNA or RNA. In some embodiments, the method includes calculating, for

each of the possible ratios of DNA or RNA in the partition and for each hypothesis, the likelihood that the hypothesis is correct by comparing either (i) the calculated allele ratios to the predicted allele ratios, or (ii) a test statistic calculated using the calculated allele ratios and the possible ratio of DNA or RNA to the expected distribution of the test statistic calculated using the predicted allele ratios and the possible ratio of DNA or RNA. In some embodiments, the combined probability for each hypothesis is determined by combining the probabilities of that hypothesis for each of the possible ratios in the partition; and the hypothesis with the greatest combined probability is selected. In some embodiments, the combined probability for each hypothesis is determined by weighting the probability of a hypothesis for a particular possible ratio based on the likelihood that the possible ratio is the correct ratio.

[215] In some embodiments, a technique selected from the group consisting of maximum likelihood estimation, maximum a-posteriori estimation, Bayesian estimation, dynamic estimation (such as dynamic Bayesian estimation), and expectation-maximization estimation is used to estimate the ratio of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample. In some embodiments, the ratio of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample is assumed to be the same for two or more (or all) of the CNVs of interest. In some embodiments, the ratio of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample is calculated for each CNV of interest.

### ***Exemplary Methods for Using Imperfectly Phased Data***

[216] In some embodiments, imperfectly phased data is used. For example, it may not be known with 100% certainty which allele is present for one or more of the loci on the first and/or second homologous chromosome segment. In some embodiments, the priors for possible haplotypes of the individual (such as haplotypes based on population based haplotype frequencies) are used in calculating the probability of each hypothesis. In some embodiments, the priors for possible haplotypes are adjusted by either using another method to phase the genetic data or by using phased data from other subjects (such as prior subjects) to refine population data used for informatics based phasing of the individual.

[217] In some embodiments, the phased genetic data comprises probabilistic data for two or more possible sets of phased genetic data, wherein each possible set of phased data comprises a possible identity of the allele present at each locus in the set of polymorphic loci on the first homologous chromosome segment and a possible identity of the allele present at each locus in the set of polymorphic loci on the

each of the possible ratios of DNA or RNA in the partition and for each hypothesis, the likelihood that the hypothesis is correct by comparing either (i) the calculated allele ratios to the predicted allele ratios, or (ii) a test statistic calculated using the calculated allele ratios and the possible ratio of DNA or RNA to the expected distribution of the test statistic calculated using the predicted allele ratios and the possible ratio of DNA or RNA. In some embodiments, the combined probability for each hypothesis is determined by combining the probabilities of that hypothesis for each of the possible ratios in the partition; and the hypothesis with the greatest combined probability is selected. In some embodiments, the combined probability for each hypothesis is determined by weighting the probability of a hypothesis for a particular possible ratio based on the likelihood that the possible ratio is the correct ratio.

[215] In some embodiments, a technique selected from the group consisting of maximum likelihood estimation, maximum a-posteriori estimation, Bayesian estimation, dynamic estimation (such as dynamic Bayesian estimation), and expectation-maximization estimation is used to estimate the ratio of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample. In some embodiments, the ratio of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample is assumed to be the same for two or more (or all) of the CNVs of interest. In some embodiments, the ratio of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample is calculated for each CNV of interest.

### ***Exemplary Methods for Using Imperfectly Phased Data***

[216] In some embodiments, imperfectly phased data is used. For example, it may not be known with 100% certainty which allele is present for one or more of the loci on the first and/or second homologous chromosome segment. In some embodiments, the priors for possible haplotypes of the individual (such as haplotypes based on population based haplotype frequencies) are used in calculating the probability of each hypothesis. In some embodiments, the priors for possible haplotypes are adjusted by either using another method to phase the genetic data or by using phased data from other subjects (such as prior subjects) to refine population data used for informatics based phasing of the individual.

[217] In some embodiments, the phased genetic data comprises probabilistic data for two or more possible sets of phased genetic data, wherein each possible set of phased data comprises a possible identity of the allele present at each locus in the set of polymorphic loci on the first homologous chromosome segment and a possible identity of the allele present at each locus in the set of polymorphic loci on the

second homologous chromosome segment. In some embodiments, the probability for at least one of the hypotheses is determined for each of the possible sets of phased genetic data. In some embodiments, the combined probability for the hypothesis is determined by combining the probabilities of the hypothesis for each of the possible sets of phased genetic data; and the hypothesis with the greatest combined probability is selected.

[218] Any of the methods disclosed herein or any known method may be used to generate imperfectly phased data (such as using population based haplotype frequencies to infer the most likely phase) for use in the claimed methods. In some embodiments, phased data is obtained by probabilistically combining haplotypes of smaller segments. For example, possible haplotypes can be determined based on possible combinations of one haplotype from a first region with another haplotype from another region from the same chromosome. The probability that particular haplotypes from different regions are part of the same, larger haplotype block on the same chromosome can be determined using, *e.g.*, population based haplotype frequencies and/or known recombination rates between the different regions.

[219] In some embodiments, a single hypothesis rejection test is used for the null hypothesis of disomy. In some embodiments, the probability of the disomy hypothesis is calculated, and the hypothesis of disomy is rejected if the probability is below a given threshold value (such as less than 1 in 1,000). If the null hypothesis is rejected, this could be due to errors in the imperfectly phased data or due to the presence of a CNV. In some embodiments, more accurate phased data is obtained (such as phased data from any of the molecular phasing methods disclosed herein to obtain actual phased data rather than bioinformatics-based inferred phased data). In some embodiments, the probability of the disomy hypothesis is recalculated using the more accurate phased data to determine if the disomy hypothesis should still be rejected. Rejection of this hypothesis indicates that a duplication or deletion of the chromosome segment is present. If desired, the false positive rate can be altered by adjusting the threshold value.

### ***Exemplary Test Statistic for Analysis of Phased Data***

[220] An exemplary test statistic is described below for analysis of phased data from a sample known or suspected of being a mixed sample containing DNA or RNA that originated from two or more cells that are not genetically identical. Let  $f$  denote the fraction of DNA or RNA of interest, for example the fraction of DNA or RNA with a CNV of interest, or the fraction of DNA or RNA from cells of interest, such as cancer cells. In some embodiments for prenatal testing,  $f$  denotes the fraction of fetal DNA, RNA, or cells in a mixture of fetal and maternal DNA, RNA, or cells. Note that this refers to the fraction of

second homologous chromosome segment. In some embodiments, the probability for at least one of the hypotheses is determined for each of the possible sets of phased genetic data. In some embodiments, the combined probability for the hypothesis is determined by combining the probabilities of the hypothesis for each of the possible sets of phased genetic data; and the hypothesis with the greatest combined probability is selected.

[218] Any of the methods disclosed herein or any known method may be used to generate imperfectly phased data (such as using population based haplotype frequencies to infer the most likely phase) for use in the claimed methods. In some embodiments, phased data is obtained by probabilistically combining haplotypes of smaller segments. For example, possible haplotypes can be determined based on possible combinations of one haplotype from a first region with another haplotype from another region from the same chromosome. The probability that particular haplotypes from different regions are part of the same, larger haplotype block on the same chromosome can be determined using, *e.g.*, population based haplotype frequencies and/or known recombination rates between the different regions.

[219] In some embodiments, a single hypothesis rejection test is used for the null hypothesis of disomy. In some embodiments, the probability of the disomy hypothesis is calculated, and the hypothesis of disomy is rejected if the probability is below a given threshold value (such as less than 1 in 1,000). If the null hypothesis is rejected, this could be due to errors in the imperfectly phased data or due to the presence of a CNV. In some embodiments, more accurate phased data is obtained (such as phased data from any of the molecular phasing methods disclosed herein to obtain actual phased data rather than bioinformatics-based inferred phased data). In some embodiments, the probability of the disomy hypothesis is recalculated using the more accurate phased data to determine if the disomy hypothesis should still be rejected. Rejection of this hypothesis indicates that a duplication or deletion of the chromosome segment is present. If desired, the false positive rate can be altered by adjusting the threshold value.

### ***Exemplary Test Statistic for Analysis of Phased Data***

[220] An exemplary test statistic is described below for analysis of phased data from a sample known or suspected of being a mixed sample containing DNA or RNA that originated from two or more cells that are not genetically identical. Let  $f$  denote the fraction of DNA or RNA of interest, for example the fraction of DNA or RNA with a CNV of interest, or the fraction of DNA or RNA from cells of interest, such as cancer cells. In some embodiments for prenatal testing,  $f$  denotes the fraction of fetal DNA, RNA, or cells in a mixture of fetal and maternal DNA, RNA, or cells. Note that this refers to the fraction of

DNA from cells of interest assuming two copies of DNA are given by each cell of interest. This differs from the DNA fraction from cells of interest at a segment that is deleted or duplicated.

DNA from cells of interest assuming two copies of DNA are given by each cell of interest. This differs from the DNA fraction from cells of interest at a segment that is deleted or duplicated.

[00221] The possible allelic values of each SNP are denoted A and B. AA, AB, BA, and BB are used to denote all possible ordered allele pairs. In some embodiments, SNPs with ordered alleles The allele ratio  $R_i$  is defined:

$$A_t$$

Let  $T$  denote the number of SNPs targeted.

[222] Without loss of generality, some embodiments focus on a single chromosome segment. In some such embodiments, all of the target SNPs are contained in the segment chromosome of interest. In other embodiments, multiple chromosome segments are analyzed for possible copy number variations.

### **MAP Estimation**

[223] This method leverages the knowledge of phasing via ordered alleles to detect the deletion or duplication of the target segment. For each SNP  $i$ , define

[224] Then define

The distributions of the  $X_t$  and  $S$  under various copy number hypotheses (such as hypotheses for disomy, deletion of the first or second homolog, or duplication of the first or second homolog) are described below.

[00221] The possible allelic values of each SNP are denoted A and B. AA, AB, BA, and BB are used to denote all possible ordered allele pairs. In some embodiments, SNPs with ordered alleles The allele ratio  $R_i$  is defined:

$$A_t$$

Let  $T$  denote the number of SNPs targeted.

[222] Without loss of generality, some embodiments focus on a single chromosome segment. In some such embodiments, all of the target SNPs are contained in the segment chromosome of interest. In other embodiments, multiple chromosome segments are analyzed for possible copy number variations.

### **MAP Estimation**

[223] This method leverages the knowledge of phasing via ordered alleles to detect the deletion or duplication of the target segment. For each SNP  $i$ , define

[224] Then define

The distributions of the  $X_t$  and  $S$  under various copy number hypotheses (such as hypotheses for disomy, deletion of the first or second homolog, or duplication of the first or second homolog) are described below.

### Disomy Hypothesis

[225] Under the hypothesis that the target segment is not deleted or duplicated,

$$p(b, n) = \Pr \{ X \sim \text{Bino}(b, n) > \frac{n}{2} \}$$

where

$$p(b, n) = \Pr \{ X \sim \text{Bino}(b, n) > \frac{n}{2} \}$$

If we assume a constant depth of read  $N$ , this gives us a Binomial distribution  $S$  with parameters

$$p(j, iv) \text{ and } T.$$

### Deletion Hypotheses

[226] Under the hypothesis that the first homo log is deleted (*i.e.*, an AB SNP becomes B, and a BA SNP becomes A), then  $R_j$  has a Binomial distribution with parameters 1 — and T for AB SNPs, and and T for BA SNPs. Therefore,

$$p(b, n) = \Pr \{ X \sim \text{Bino}(b, n) > \frac{n}{2} \}$$

If we assume a constant depth of read  $N$ , this gives a Binomial distribution  $S$  with parameters

$$p(jr^w) \text{ and } T.$$

[227] Under the hypothesis that the second homo log is deleted (*i.e.*, an AB SNP becomes A, and a BA SNP becomes B), then  $R_i$  has a Binomial distribution with parameters and T for AB SNPs, and 1 — and T for BA SNPs. Therefore,

### Disomy Hypothesis

[225] Under the hypothesis that the target segment is not deleted or duplicated,

$$p(b, n) = \Pr \{ X \sim \text{Bino}(b, n) > \frac{n}{2} \mid \mathcal{H} \}$$

where

$$p(b, n) = \Pr \{ X \sim \text{Bino}(b, n) > \frac{n}{2} \mid \mathcal{H} \}$$

If we assume a constant depth of read  $N$ , this gives us a Binomial distribution  $S$  with parameters

$$p(j, n) \text{ and } T.$$

### Deletion Hypotheses

[226] Under the hypothesis that the first homo log is deleted (*i.e.*, an AB SNP becomes B, and a BA SNP becomes A), then  $R_j$  has a Binomial distribution with parameters  $1 - \frac{1}{2}$  and  $T$  for AB SNPs, and  $\frac{1}{2}$  and  $T$  for BA SNPs. Therefore,

$$p(b, n) = \Pr \{ X \sim \text{Bino}(b, n) > \frac{n}{2} \mid \mathcal{H} \}$$

If we assume a constant depth of read  $N$ , this gives a Binomial distribution  $S$  with parameters

$$p(j, n) \text{ and } T.$$

[227] Under the hypothesis that the second homo log is deleted (*i.e.*, an AB SNP becomes A, and a BA SNP becomes B), then  $R_i$  has a Binomial distribution with parameters  $1 - \frac{1}{2}$  and  $T$  for AB SNPs, and  $\frac{1}{2}$  and  $T$  for BA SNPs. Therefore,

$$P(X_i = j) = \binom{N}{j} p^j (1-p)^{N-j}$$

If we assume a constant depth of read  $N$ , this gives a Binomial distribution  $S$  with parameters

$$p = \frac{1}{2} \text{ and } T = \frac{1}{2}$$

### Duplication Hypotheses

[228] Under the hypothesis that the first homo log is duplicated (*i.e.*, an AB SNP becomes AAB, and a BA SNP becomes BBA), then has a Binomial distribution with parameters  $p = \frac{1}{2}$  and  $T = \frac{1}{2}$  for AB SNPs, and  $p = \frac{1}{2}$  and  $T = \frac{1}{2}$  for BA SNPs. Therefore,

$$P(X_i = j) = \binom{N}{j} p^j (1-p)^{N-j}$$

If we assume a constant depth of read  $N$ , this gives us a Binomial distribution  $S$  with parameters

$$p = \frac{1}{2} \text{ and } T = \frac{1}{2}$$

[229] Under the hypothesis that the second homo log is duplicated (*i.e.*, an AB SNP becomes ABB, and a BA SNP becomes BAA), then has a Binomial distribution with parameters

$p = \frac{1}{2}$  and  $T = \frac{1}{2}$  for AB SNPs, and  $p = \frac{1}{2}$  and  $T = \frac{1}{2}$  for BA SNPs. Therefore,

$$P(X_i = j) = \binom{N}{j} p^j (1-p)^{N-j}$$

$$p = \frac{1}{2} \text{ and } T = \frac{1}{2}$$

If we assume a constant depth of read  $N$ , this gives a Binomial distribution  $S$  with parameters

$$p = \frac{1}{2} \text{ and } T = \frac{1}{2}$$

$$P(X_i = j) = \binom{N}{j} p^j (1-p)^{N-j}$$

If we assume a constant depth of read  $N$ , this gives a Binomial distribution  $S$  with parameters

$$p = \frac{1}{2} \text{ and } T = \frac{1}{2}$$

### Duplication Hypotheses

[228] Under the hypothesis that the first homo log is duplicated (*i.e.*, an AB SNP becomes AAB, and a BA SNP becomes BBA), then has a Binomial distribution with parameters  $p = \frac{1}{2}$  and  $T = \frac{1}{2}$  for AB SNPs, and  $p = \frac{1}{2}$  and  $T = \frac{1}{2}$  for BA SNPs. Therefore,

$$P(X_i = j) = \binom{N}{j} p^j (1-p)^{N-j}$$

If we assume a constant depth of read  $N$ , this gives us a Binomial distribution  $S$  with parameters

$$p = \frac{1}{2} \text{ and } T = \frac{1}{2}$$

[229] Under the hypothesis that the second homo log is duplicated (*i.e.*, an AB SNP becomes ABB, and a BA SNP becomes BAA), then has a Binomial distribution with parameters

$p = \frac{1}{2}$  and  $T = \frac{1}{2}$  for AB SNPs, and  $p = \frac{1}{2}$  and  $T = \frac{1}{2}$  for BA SNPs. Therefore,

$$P(X_i = j) = \binom{N}{j} p^j (1-p)^{N-j}$$

$$p = \frac{1}{2} \text{ and } T = \frac{1}{2}$$

If we assume a constant depth of read  $N$ , this gives a Binomial distribution  $S$  with parameters

$$p = \frac{1}{2} \text{ and } T = \frac{1}{2}$$

## Classification

[230] As demonstrated in the sections above, is a binary random variable with

$$p(\wedge, N^j \text{ given disomy} \\ p \text{ homolog } \pm \text{ deletion} \\ Pr\{X_i = 1\} = i \cdot 1 - p(jlj'Ni) \text{ homolog } 2 \text{ deletion } \pm - p(\text{---} \\ 7, iVj) \text{ homolog } 1 \text{ duplication} \\ r \setminus 2 + / \text{ or} \\ \text{homolog } 2 \text{ duplication}$$

[231] This allows one to calculate the probability of the test statistic  $S$  under each hypothesis. The probability of each hypothesis given the measured data can be calculated. In some embodiments, the hypothesis with the greatest probability is selected. If desired, the distribution

on  $S$  can be simplified by either approximating each  $N_i$  with a constant depth of reach  $N$  or by truncating the depth of reads to a constant  $N$ . This simplification gives

## given disomy

[232] The value for  $f$  can be estimate by selecting the most likely value of  $f$  given the measured data, such as the value of  $f$  that generates the best data fit using an algorithm (*e.g.*, a search algorithm) such as maximum likelihood estimation, maximum a-posteriori estimation, or Bayesian estimation. In some embodiments, multiple chromosome segments are analyzed and a value for  $f$  is estimated based on the data for each segment. If all the target cells have these duplications or deletions, the estimated values for  $f$ /based on data for these different segments are similar. In some embodiments,  $f$  is experimentally measured such as by determining the fraction of DNA or RNA from cancer cells based on methylation differences (hypomethylation or hypermethylation) between cancer and non-cancerous DNA or RNA.

[233] In some embodiments for mixed samples of fetal and maternal nucleic acids, the value of  $f$  is the fetal fraction, that is the fraction of fetal DNA (or RNA) out of the total amount of DNA (or RNA) in the sample. In some embodiments, the fetal fraction is determined by obtaining genotypic data from a maternal blood sample (or fraction thereof) for a set of polymorphic loci on at least one chromosome that is expected to be disomic in both the mother and the fetus; creating a plurality of hypotheses each corresponding to different possible fetal fractions at the chromosome; building a model for the expected allele measurements in the blood sample at the set of polymorphic loci on the chromosome for possible

## Classification

[230] As demonstrated in the sections above, is a binary random variable with

$$p(\wedge, N^j \text{ given disomy} \\ p \text{ homolog } \pm \text{ deletion} \\ Pr\{X_i = 1\} = i \cdot 1 - p(jlj'Ni) \text{ homolog } 2 \text{ deletion } \pm - p(\text{---} \\ 7, iVj) \text{ homolog } 1 \text{ duplication} \\ r \setminus 2 + / \text{ or} \\ \text{homolog } 2 \text{ duplication}$$

[231] This allows one to calculate the probability of the test statistic  $S$  under each hypothesis. The probability of each hypothesis given the measured data can be calculated. In some embodiments, the hypothesis with the greatest probability is selected. If desired, the distribution

on  $S$  can be simplified by either approximating each  $N_i$  with a constant depth of reach  $N$  or by truncating the depth of reads to a constant  $N$ . This simplification gives

*given disomy*

[232] The value for  $/$  can be estimate by selecting the most likely value of  $/$  given the measured data, such as the value of  $/$  that generates the best data fit using an algorithm (*e.g.*, a search algorithm) such as maximum likelihood estimation, maximum a-posteriori estimation, or Bayesian estimation. In some embodiments, multiple chromosome segments are analyzed and a value for  $/$  is estimated based on the data for each segment. If all the target cells have these duplications or deletions, the estimated values for  $/$  based on data for these different segments are similar. In some embodiments,  $/$  is experimentally measured such as by determining the fraction of DNA or RNA from cancer cells based on methylation differences (hypomethylation or hypermethylation) between cancer and non-cancerous DNA or RNA.

[233] In some embodiments for mixed samples of fetal and maternal nucleic acids, the value of  $/$  is the fetal fraction, that is the fraction of fetal DNA (or RNA) out of the total amount of DNA (or RNA) in the sample. In some embodiments, the fetal fraction is determined by obtaining genotypic data from a maternal blood sample (or fraction thereof) for a set of polymorphic loci on at least one chromosome that is expected to be disomic in both the mother and the fetus; creating a plurality of hypotheses each corresponding to different possible fetal fractions at the chromosome; building a model for the expected allele measurements in the blood sample at the set of polymorphic loci on the chromosome for possible

fetal fractions; calculating a relative probability of each of the fetal fractions hypotheses using the model and the allele measurements from the blood sample or fraction thereof; and determining the fetal fraction in the blood sample by selecting the fetal fraction corresponding to the hypothesis with the greatest probability. In some embodiments, the fetal fraction is determined by identifying those polymorphic loci where the mother is homozygous for a first allele at the polymorphic locus, and the father is (i) heterozygous for the first allele and a second allele or (ii) homozygous for a second allele at the polymorphic locus; and using the amount of the second allele detected in the blood sample for each of the identified polymorphic loci to determine the fetal fraction in the blood sample (see, *e.g.*, US Publ. No. 2012/0185176, filed March 29, 2012, and US Pub. No. 2014/0065621, filed March 13, 2013 which are each incorporated herein by reference in their entirety).

[234] Another method for determining fetal fraction includes using a high throughput DNA sequencer to count alleles at a large number of polymorphic (such as SNP) genetic loci and modeling the likely fetal fraction (see, for example, US Publ. No. 2012/0264121, which is incorporated herein by reference in its entirety). Another method for calculating fetal fraction can be found in Sparks *et al* "Noninvasive prenatal detection and selective analysis of cell-free DNA obtained from maternal blood: evaluation for trisomy 21 and trisomy 18," Am J Obstet Gynecol 2012;206:319.e1-9, which is incorporated herein by reference in its entirety. In some embodiments, fetal fraction is determined using a methylation assay (*see, e.g.*, US Patent Nos. 7,754,428; 7,901,884; and 8,166,382, which are each incorporated herein by reference in their entirety) that assumes certain loci are methylated or preferentially methylated in the fetus, and those same loci are unmethylated or preferentially unmethylated in the mother.

[235] FIGs. 1A-13D are graphs showing the distribution of the test statistic  $S$  divided by  $T$  (the number of SNPs) (" $S/T$ ") for various copy number hypotheses for various depth of reads and tumor fractions (where  $f$  is the fraction of tumor DNA out of total DNA) for an increasing number of SNPs.

#### *Single Hypothesis Rejection*

[236] The distribution of  $S$  for the disomy hypothesis does not depend on  $f$ . Thus, the probability of the measured data can be calculated for the disomy hypothesis without calculating  $f$ . A single hypothesis rejection test can be used for the null hypothesis of disomy. In some embodiments, the probability of  $S$  under the disomy hypothesis is calculated, and the hypothesis of disomy is rejected if the probability is below a given threshold value (such as less than  $\alpha$  in 1,000). This indicates that a duplication or deletion

fetal fractions; calculating a relative probability of each of the fetal fractions hypotheses using the model and the allele measurements from the blood sample or fraction thereof; and determining the fetal fraction in the blood sample by selecting the fetal fraction corresponding to the hypothesis with the greatest probability. In some embodiments, the fetal fraction is determined by identifying those polymorphic loci where the mother is homozygous for a first allele at the polymorphic locus, and the father is (i) heterozygous for the first allele and a second allele or (ii) homozygous for a second allele at the polymorphic locus; and using the amount of the second allele detected in the blood sample for each of the identified polymorphic loci to determine the fetal fraction in the blood sample (see, *e.g.*, US Publ. No. 2012/0185176, filed March 29, 2012, and US Pub. No. 2014/0065621, filed March 13, 2013 which are each incorporated herein by reference in their entirety).

[234] Another method for determining fetal fraction includes using a high throughput DNA sequencer to count alleles at a large number of polymorphic (such as SNP) genetic loci and modeling the likely fetal fraction (see, for example, US Publ. No. 2012/0264121, which is incorporated herein by reference in its entirety). Another method for calculating fetal fraction can be found in Sparks *et al* "Noninvasive prenatal detection and selective analysis of cell-free DNA obtained from maternal blood: evaluation for trisomy 21 and trisomy 18," Am J Obstet Gynecol 2012;206:319.e1-9, which is incorporated herein by reference in its entirety. In some embodiments, fetal fraction is determined using a methylation assay (*see, e.g.*, US Patent Nos. 7,754,428; 7,901,884; and 8,166,382, which are each incorporated herein by reference in their entirety) that assumes certain loci are methylated or preferentially methylated in the fetus, and those same loci are unmethylated or preferentially unmethylated in the mother.

[235] FIGs. 1A-13D are graphs showing the distribution of the test statistic  $S$  divided by  $T$  (the number of SNPs) (" $S/T$ ") for various copy number hypotheses for various depth of reads and tumor fractions (where  $f$  is the fraction of tumor DNA out of total DNA) for an increasing number of SNPs.

### *Single Hypothesis Rejection*

[236] The distribution of  $S$  for the disomy hypothesis does not depend on  $f$ . Thus, the probability of the measured data can be calculated for the disomy hypothesis without calculating  $f$ . A single hypothesis rejection test can be used for the null hypothesis of disomy. In some embodiments, the probability of  $S$  under the disomy hypothesis is calculated, and the hypothesis of disomy is rejected if the probability is below a given threshold value (such as less than  $\alpha$  in 1,000). This indicates that a duplication or deletion

of the chromosome segment is present. If desired, the false positive rate can be altered by adjusting the threshold value.

### *Exemplary Methods for Analysis of Phased Data*

[237] Exemplary methods are described below for analysis of data from a sample known or suspected of being a mixed sample containing DNA or RNA that originated from two or more cells that are not genetically identical. In some embodiments, phased data is used. In some embodiments, the method involves determining, for each calculated allele ratio, whether the calculated allele ratio is above or below the expected allele ratio and the magnitude of the difference for a particular locus. In some embodiments, a likelihood distribution is determined for the allele ratio at a locus for a particular hypothesis and the closer the calculated allele ratio is to the center of the likelihood distribution, the more likely the hypothesis is correct. In some embodiments, the method involves determining the likelihood that a hypothesis is correct for each locus. In some embodiments, the method involves determining the likelihood that a hypothesis is correct for each locus, and combining the probabilities of that hypothesis for each locus, and the hypothesis with the greatest combined probability is selected. In some embodiments, the method involves determining the likelihood that a hypothesis is correct for each locus and for each possible ratio of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample. In some embodiments, a combined probability for each hypothesis is determined by combining the probabilities of that hypothesis for each locus and each possible ratio, and the hypothesis with the greatest combined probability is selected.

[238] In one embodiment, the following hypotheses are considered:  $H_{00}$  (all cells are normal),  $H_{10}$  (presence of cells with only homolog 1, hence homolog 2 deletion),  $H_{01}$  (presence of cells with only homolog 2, hence homolog 1 deletion),  $H_{11}$  (presence of cells with homolog 1 duplication),  $H_{22}$  (presence of cells with homolog 2 duplication). For a fraction  $f$  of target cells such as cancer cells or mosaic cells (or the fraction of DNA or RNA from the target cells), the expected allele ratio for heterozygous ( $AB$  or  $BA$ ) SNPs can be found as follows:

$$r\{AB, H_n\} = r\{BA, H_n\} = r\{AB, H_{10}\} = r\{BA, H_{01}\} = 0.5,$$

$$r\{AB, H_{01}\} = r\{BA, H_{10}\} = \frac{2-f}{2+f}$$

of the chromosome segment is present. If desired, the false positive rate can be altered by adjusting the threshold value.

### *Exemplary Methods for Analysis of Phased Data*

[237] Exemplary methods are described below for analysis of data from a sample known or suspected of being a mixed sample containing DNA or RNA that originated from two or more cells that are not genetically identical. In some embodiments, phased data is used. In some embodiments, the method involves determining, for each calculated allele ratio, whether the calculated allele ratio is above or below the expected allele ratio and the magnitude of the difference for a particular locus. In some embodiments, a likelihood distribution is determined for the allele ratio at a locus for a particular hypothesis and the closer the calculated allele ratio is to the center of the likelihood distribution, the more likely the hypothesis is correct. In some embodiments, the method involves determining the likelihood that a hypothesis is correct for each locus. In some embodiments, the method involves determining the likelihood that a hypothesis is correct for each locus, and combining the probabilities of that hypothesis for each locus, and the hypothesis with the greatest combined probability is selected. In some embodiments, the method involves determining the likelihood that a hypothesis is correct for each locus and for each possible ratio of DNA or RNA from the one or more target cells to the total DNA or RNA in the sample. In some embodiments, a combined probability for each hypothesis is determined by combining the probabilities of that hypothesis for each locus and each possible ratio, and the hypothesis with the greatest combined probability is selected.

[238] In one embodiment, the following hypotheses are considered:  $H_{00}$  (all cells are normal),  $H_{10}$  (presence of cells with only homolog 1, hence homolog 2 deletion),  $H_{01}$  (presence of cells with only homolog 2, hence homolog 1 deletion),  $H_{11}$  (presence of cells with homolog 1 duplication),  $H_{22}$  (presence of cells with homolog 2 duplication). For a fraction  $f$  of target cells such as cancer cells or mosaic cells (or the fraction of DNA or RNA from the target cells), the expected allele ratio for heterozygous ( $AB$  or  $BA$ ) SNPs can be found as follows:

$$r\{AB, H_n\} = r\{BA, H_n\} = r\{AB, H_{10}\} = r\{BA, H_{01}\} = 0.5,$$

$$r\{AB, H_{01}\} = r\{BA, H_{10}\} = \frac{2-f}{2+f}$$

$$r(AB, H_{21}) = r(BA, H_{12}) =$$

$$r(AB, H_n) = r(BA, H_{21}) = 2 + \frac{1}{n}$$

Bias, Contamination, and Sequencing Error Correction (*fn v*)

$$r(AB, H_{21}) = r(BA, H_{12}) =$$

$$r(AB, H_n) = r(BA, H_{21}) = \frac{2 + \frac{1}{n}}{2 + \frac{1}{n} + \frac{1}{n}} = \frac{2 + \frac{1}{n}}{2 + \frac{2}{n}} = \frac{2n + 1}{2n + 2}$$

Bias, Contamination, and Sequencing Error Correction (*fn v*)

$s_e$  denote the sequencing errors for calling one allele a different allele (such as by erroneously detecting an A allele when a B allele is present).

[241] One can find the observed allele ratio  $q$  for a given expected allele ratio  $r$  as follows:

$$q(r, c_a, r_{c_{ts}}, c_{ts}, r_{c_t}) = (r(1 - s_e^{r''})) + (1 - r)s_e^{r''}(1 - c_{ts} - c_g) + (r_{c_{ts}}(1 - s_e^{r''})) + (1 - r_{c_{ts}})s_e^{r''}$$

[242] Since the contaminant genotypes are unknown, population frequencies can be used to find  $P(r_{c_g})$ . More specifically, let  $p$  be the population frequency for one of the alleles (which may be referred to as a reference allele). Then, we have  $P(r_c = 0) = (1 - p)^2$ ,  $P(r_{c_g} = 0.5) = 2p(1 - p)$ , and  $P(r_{c_g} = 1) = p^2$ . The conditional expectation over  $r_{c_g}$  in the equation above can be used to determine the  $E[q(r, c_a, r_{c_n}, c_{ts}, r_{c_g})]$ . Note that the ambient and genotyped contamination are determined using the homozygous SNPs, hence they are not affected by the absence or presence of deletions or duplications. Moreover, it is possible to measure the ambient and genotyped contamination using a reference chromosome if desired.

#### Likelihood at Each SNP

[243] The equation below gives the probability of observing  $n_A$  and  $n_B$  given an expected allele ratio  $r$ :

$$P\{n_A, n_B | r\} = \binom{n_A + n_B}{n_A} r^{n_A} (1 - r)^{n_B} \quad (2)$$

[244] Let  $I_s$  denote the data for SNP  $s$ . For each hypothesis  $IEH = \{i | i \in \{0, 1, \dots, M\}\}$ , one can let  $r = r(A_s, h)$  or  $r = r(B_s, h)$  in the equation (1) and find the conditional expectation over  $r_{c_g}$  to determine the observed allele ratio  $E[q(r, c_a, r_{c_a}, c_g, r_{c_g})]$ . Then, letting  $r = E[q(r, c_a, r_{c_a}, c_g, r_{c_g})]$  in equation (2) one can determine  $P(D_s | h, f)$ .

#### Search Algorithm

the variability in the allele ratios may be high, hence this ensures that SNPs will not be trimmed due to mosaicism.

[245] Let  $F = \{f_1, \dots, f_{i_f}\}$  denote the search space for the mosaicism percentage (such as the tumor fraction). One can determine  $P(D_s | h, f)$  at each SNP  $s$  and  $f \in F$ , and combine the

$s_e$  denote the sequencing errors for calling one allele a different allele (such as by erroneously detecting an A allele when a B allele is present).

[241] One can find the observed allele ratio  $q$  for a given expected allele ratio  $r$  as follows:

$$q(r, c_a, r_{c_{ts}}, c_{ts}, r_{c_l}) = (r(1 - s_e^{r''})) + (1 - r)4^{r''})(1 - c_{ts} - c_g) + (r_{c_{ts}}(1 - s_e^{r''})) + (1 - r_{c_{ts}})s_e^{r''}$$

[242] Since the contaminant genotypes are unknown, population frequencies can be used to find  $P\{r_{c_g}\}$ . More specifically, let  $p$  be the population frequency for one of the alleles (which may be referred to as a reference allele). Then, we have  $P(r_c = 0) = (1 - p)^2$ ,  $P(r_{c_g} = 0.5) = 2p(1 - p)$ , and  $P(r_{c_g} = 1) = p^2$ . The conditional expectation over  $r_{c_g}$  in the equation above can be used to determine the  $E[q(r, c_a, r_{c_n}, c_{ts}, r_{c_l})]$ . Note that the ambient and genotyped contamination are determined using the homozygous SNPs, hence they are not affected by the absence or presence of deletions or duplications. Moreover, it is possible to measure the ambient and genotyped contamination using a reference chromosome if desired.

#### Likelihood at Each SNP

[243] The equation below gives the probability of observing  $n_A$  and  $n_B$  given an expected allele ratio  $r$ :

$$P\{n_A, n_B | r\} = \text{binom}\{n_A, n_B, r\} = r^{n_A} (1 - r)^{n_B}. \quad (2)$$

[244] Let  $I_s$  denote the data for SNP  $s$ . For each hypothesis  $IEH = \{i | i \in \{0, 1, 2\}\}$ , one can let  $r = r(A5, h)$  or  $r = r(BA, h)$  in the equation (1) and find the conditional expectation over  $r_{c_g}$  to determine the observed allele ratio  $E[q(r, c_a, r_{c_a}, c_g, r_{c_g})]$ . Then, letting  $r = E[q(r, c_a, r_{c_a}, c_g, r_{c_g})]$  in equation (2) one can determine  $P(D_s | h, f)$ .

#### Search Algorithm

the variability in the allele ratios may be high, hence this ensures that SNPs will not be trimmed due to mosaicism.

[245] Let  $F = \{f_1, \dots, f_{i_f}\}$  denote the search space for the mosaicism percentage (such as the tumor fraction). One can determine  $P(D_s | h, f)$  at each SNP  $s$  and  $f \in F$ , and combine the

log likelihood over all SNPs.

[246] The algorithm goes over each  $f$  for each hypothesis. Using a search method, one concludes that mosaicism exists if there is a range  $F^*$  of  $f$  where the confidence of the deletion or duplication hypothesis is higher than the confidence of the no deletion and no duplication hypotheses. In some embodiments, the maximum likelihood estimate for  $P(D_s|h,f)$  in  $F^*$  is determined.

If desired, the conditional expectation over  $f \in F^*$  may be determined. If desired, the confidence for each hypothesis can be determined.

#### Additional Embodiments

[247] In some embodiments, a beta binomial distribution is used instead of binomial distribution. In some embodiments, the parameters of the beta binomial are estimated using the expected bias and bias variance (over all cycles). In some embodiments, a reference chromosome or chromosome segment is used to determine the sample specific parameters of beta binomial.

#### Theoretical Performance using Simulations

[248] If desired, one can evaluate the theoretical performance of the algorithm by randomly assigning number of reference reads to a SNP with given depth of read (DOR). For the normal case, use  $p = 0.5$  for the binomial probability parameter, and for deletions or duplications,  $p$  is revised accordingly. Exemplary input parameters for each simulation are as follows: (1) number of snps  $S$ , (2) constant DOR  $D$  per snp, (3)  $p$ , and (4) number of experiments.

#### [249] First Simulation Experiment

[250] setting (hence 24,000 experiments with phase, and 24,000 without phase). We simulated the number of reads from a binomial distribution (if desired, other distributions can be used). The false positive rate (in the case of  $p = 0\%$ ) and false negative rate (in the case of  $p > 0\%$ ) were determined both with or without phase information. False positive rates are listed in Figure 26. Note that phase information is very helpful, especially for  $S = 1000$ ,  $D = 1000$ . Although for  $S = 500$ ,  $D = 500$ , the algorithm has the highest false positive rates with or without

log likelihood over all SNPs.

[246] The algorithm goes over each  $f$  for each hypothesis. Using a search method, one concludes that mosaicism exists if there is a range  $F^*$  of  $f$  where the confidence of the deletion or duplication hypothesis is higher than the confidence of the no deletion and no duplication hypotheses. In some embodiments, the maximum likelihood estimate for  $P(D_s|h,f)$  in  $F^*$  is determined.

If desired, the conditional expectation over  $f \in F^*$  may be determined. If desired, the confidence for each hypothesis can be determined.

#### Additional Embodiments

[247] In some embodiments, a beta binomial distribution is used instead of binomial distribution. In some embodiments, the parameters of the beta binomial are estimated using the expected bias and bias variance (over all cycles). In some embodiments, a reference chromosome or chromosome segment is used to determine the sample specific parameters of beta binomial.

#### Theoretical Performance using Simulations

[248] If desired, one can evaluate the theoretical performance of the algorithm by randomly assigning number of reference reads to a SNP with given depth of read (DOR). For the normal case, use  $p = 0.5$  for the binomial probability parameter, and for deletions or duplications,  $p$  is revised accordingly. Exemplary input parameters for each simulation are as follows: (1) number of snps  $S$ , (2) constant DOR  $D$  per snp, (3)  $p$ , and (4) number of experiments.

#### [249] First Simulation Experiment

[250] setting (hence 24,000 experiments with phase, and 24,000 without phase). We simulated the number of reads from a binomial distribution (if desired, other distributions can be used). The false positive rate (in the case of  $p = 0\%$ ) and false negative rate (in the case of  $p > 0\%$ ) were determined both with or without phase information. False positive rates are listed in Figure 26. Note that phase information is very helpful, especially for  $S = 1000$ ,  $D = 1000$ . Although for  $S = 500$ ,  $D = 500$ , the algorithm has the highest false positive rates with or without

phase out of the conditions tested.

False negative rates are listed in Figure 27.

[251] Phase information is particularly useful for low mosaicism percentages ( $< 3\%$ ). Without phase information, a high level of false negatives were observed for  $p = 1\%$  because the confidence on deletion is determined by assigning equal chance to  $H \setminus Q$  and Mu, and a small deviation in favor of one hypothesis is not sufficient to compensate for the low likelihood from the other hypothesis. This applies to duplications as well. Note also that the algorithm seems to be more sensitive to depth of read compared to number of SNPs. For the results with phase information, we assume that perfect phase information is available for a high number of consecutive heterozygous SNPs. If desired, haplotype information can be obtained by probabilistically combining haplotypes on smaller segments.

#### Second Simulation Experiment

##### [252] Exemplary Methods for Detecting Deletions and Duplications Without Phased Data

[253] In some embodiments, unphased genetic data is used to determine if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of an individual (such as in the genome of one or more cells or in cfDNA or cfRNA). In some embodiments, phased genetic data is used but the phasing is ignored. In some embodiments, the sample of DNA or RNA is a mixed sample of cfDNA or cfRNA from the individual that includes cfDNA or cfRNA from two or more genetically different cells. In some embodiments, the method utilizes the magnitude of the difference between the calculated allele ratio and the expected allele ratio for each of the loci.

[254] In some embodiments, the method involves obtaining genetic data at a set of polymorphic loci on the chromosome or chromosome segment in a sample of DNA or RNA from one or more cells from the individual by measuring the quantity of each allele at each locus. In some embodiments, allele ratios are calculated for the loci that are heterozygous in at least one cell from which the sample was derived (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother). In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus. In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles (such as the allele on the first homologous chromosome segment) divided by the measured quantity of one or more other alleles

phase out of the conditions tested.

False negative rates are listed in Figure 27.

[251] Phase information is particularly useful for low mosaicism percentages ( $< 3\%$ ). Without phase information, a high level of false negatives were observed for  $p = 1\%$  because the confidence on deletion is determined by assigning equal chance to  $H \setminus Q$  and Mu, and a small deviation in favor of one hypothesis is not sufficient to compensate for the low likelihood from the other hypothesis. This applies to duplications as well. Note also that the algorithm seems to be more sensitive to depth of read compared to number of SNPs. For the results with phase information, we assume that perfect phase information is available for a high number of consecutive heterozygous SNPs. If desired, haplotype information can be obtained by probabilistically combining haplotypes on smaller segments.

#### Second Simulation Experiment

##### [252] Exemplary Methods for Detecting Deletions and Duplications Without Phased Data

[253] In some embodiments, unphased genetic data is used to determine if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of an individual (such as in the genome of one or more cells or in cfDNA or cfRNA). In some embodiments, phased genetic data is used but the phasing is ignored. In some embodiments, the sample of DNA or RNA is a mixed sample of cfDNA or cfRNA from the individual that includes cfDNA or cfRNA from two or more genetically different cells. In some embodiments, the method utilizes the magnitude of the difference between the calculated allele ratio and the expected allele ratio for each of the loci.

[254] In some embodiments, the method involves obtaining genetic data at a set of polymorphic loci on the chromosome or chromosome segment in a sample of DNA or RNA from one or more cells from the individual by measuring the quantity of each allele at each locus. In some embodiments, allele ratios are calculated for the loci that are heterozygous in at least one cell from which the sample was derived (such as the loci that are heterozygous in the fetus and/or heterozygous in the mother). In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus. In some embodiments, the calculated allele ratio for a particular locus is the measured quantity of one of the alleles (such as the allele on the first homologous chromosome segment) divided by the measured quantity of one or more other alleles

(such as the allele on the second homologous chromosome segment) for the locus. The calculated allele ratios and expected allele ratios may be calculated using any of the methods described herein or any standard method (such as any mathematical transformation of the calculated allele ratios or expected allele ratios described herein).

**[255]** In some embodiments, a test statistic is calculated based on the magnitude of the difference between the calculated allele ratio and the expected allele ratio for each of the loci. In some embodiments, the test statistic  $A$  is calculated using the following formula

$A =$

wherein  $\delta_i$  is the magnitude of the difference between the calculated allele ratio and the expected allele ratio for the  $i$ th loci;

wherein  $\bar{\delta}$  is the mean value of  $\delta_i$ , and

wherein  $S$  is the standard deviation of  $\delta_i$ .

**[256]** For example, we can define  $\delta_i$  as follows when the expected allele ratio is 0.5:

Values for  $\bar{\delta}$  and  $S$  can be computed using the fact that  $R_i$  is a Binomial random variable. In some embodiments, the standard deviation is assumed to be the same for all the loci. In some embodiments, the average or weighted average value of the standard deviation or an estimate of the standard deviation is used for the value of  $S$ . In some embodiments, the test statistic is assumed to have a normal distribution. For example, the central limit theorem implies that the distribution of  $A$  converges to a standard normal as the number of loci (such as the number of SNPs  $T$ ) grows large.

**[257]** In some embodiments, a set of one or more hypotheses specifying the number of copies of the chromosome or chromosome segment in the genome of one or more of the cells are enumerated. In some embodiments, the hypothesis that is most likely based on the test statistic is selected, thereby determining the number of copies of the chromosome or chromosome segment in the genome of one or more of the cells. In some embodiments, a hypothesis is selected if the probability that the test statistic belongs to a distribution of the test statistic for that hypothesis is above an upper threshold; one or more of the hypotheses is rejected if the probability that the test statistic belongs to the distribution of the test statistic for that hypothesis is below a lower threshold; or a hypothesis is neither selected nor rejected if the probability that the test statistic belongs to the distribution of the test statistic for that hypothesis is between the lower threshold and the upper threshold, or if the probability is not determined with sufficiently high confidence. In some embodiments, an upper and/or lower threshold is determined from

(such as the allele on the second homologous chromosome segment) for the locus. The calculated allele ratios and expected allele ratios may be calculated using any of the methods described herein or any standard method (such as any mathematical transformation of the calculated allele ratios or expected allele ratios described herein).

**[255]** In some embodiments, a test statistic is calculated based on the magnitude of the difference between the calculated allele ratio and the expected allele ratio for each of the loci. In some embodiments, the test statistic  $A$  is calculated using the following formula

$A =$

wherein  $\delta_i$  is the magnitude of the difference between the calculated allele ratio and the expected allele ratio for the  $i$ th loci;

wherein  $\bar{\delta}$  is the mean value of  $\delta_i$ , and

wherein  $S$  is the standard deviation of  $\delta_i$ .

**[256]** For example, we can define  $\delta_i$  as follows when the expected allele ratio is 0.5:

Values for  $\bar{\delta}$  and  $S$  can be computed using the fact that  $R_i$  is a Binomial random variable. In some embodiments, the standard deviation is assumed to be the same for all the loci. In some embodiments, the average or weighted average value of the standard deviation or an estimate of the standard deviation is used for the value of  $S$ . In some embodiments, the test statistic is assumed to have a normal distribution. For example, the central limit theorem implies that the distribution of  $A$  converges to a standard normal as the number of loci (such as the number of SNPs  $T$ ) grows large.

**[257]** In some embodiments, a set of one or more hypotheses specifying the number of copies of the chromosome or chromosome segment in the genome of one or more of the cells are enumerated. In some embodiments, the hypothesis that is most likely based on the test statistic is selected, thereby determining the number of copies of the chromosome or chromosome segment in the genome of one or more of the cells. In some embodiments, a hypothesis is selected if the probability that the test statistic belongs to a distribution of the test statistic for that hypothesis is above an upper threshold; one or more of the hypotheses is rejected if the probability that the test statistic belongs to the distribution of the test statistic for that hypothesis is below a lower threshold; or a hypothesis is neither selected nor rejected if the probability that the test statistic belongs to the distribution of the test statistic for that hypothesis is between the lower threshold and the upper threshold, or if the probability is not determined with sufficiently high confidence. In some embodiments, an upper and/or lower threshold is determined from

an empirical distribution, such as a distribution from training data (such as samples with a known copy number, such as diploid samples or samples known to have a particular deletion or duplication). Such an empirical distribution can be used to select a threshold for a single hypothesis rejection test.

**[258]** Note that the test statistic  $A$  is independent of  $S$  and therefore both can be used independently, if desired.

#### Exemplary Methods for Detecting Deletions and Duplications Using Allele Distributions or Patterns

**[259]** This section includes methods for determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment. In some embodiments, the method involves enumerating (i) a plurality of hypotheses specifying the number of copies of the chromosome or chromosome segment that are present in the genome of one or more cells (such as cancer cells) of the individual or (ii) a plurality of hypotheses specifying the degree of overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of one or more cells of the individual. In some embodiments, the method involves obtaining genetic data from the individual at a plurality of polymorphic loci (such as SNP loci) on the chromosome or chromosome segment. In some embodiments, a probability distribution of the expected genotypes of the individual for each of the hypotheses is created. In some embodiments, a data fit between the obtained genetic data of the individual and the probability distribution of the expected genotypes of the individual is calculated. In some embodiments, one or more hypotheses are ranked according to the data fit, and the hypothesis that is ranked the highest is selected. In some embodiments, a technique or algorithm, such as a search algorithm, is used for one or more of the following steps: calculating the data fit, ranking the hypotheses, or selecting the hypothesis that is ranked the highest. In some embodiments, the data fit is a fit to a beta-binomial distribution or a fit to a binomial distribution. In some embodiments, the technique or algorithm is selected from the group consisting of maximum likelihood estimation, maximum a-posteriori estimation, Bayesian estimation, dynamic estimation (such as dynamic Bayesian estimation), and expectation-maximization estimation. In some embodiments, the method includes applying the technique or algorithm to the obtained genetic data and the expected genetic data.

**[260]** In some embodiments, the method involves enumerating (i) a plurality of hypotheses specifying the number of copies of the chromosome or chromosome segment that are present in the genome of one or more cells (such as cancer cells) of the individual or (ii) a plurality of hypotheses

an empirical distribution, such as a distribution from training data (such as samples with a known copy number, such as diploid samples or samples known to have a particular deletion or duplication). Such an empirical distribution can be used to select a threshold for a single hypothesis rejection test.

**[258]** Note that the test statistic  $A$  is independent of  $S$  and therefore both can be used independently, if desired.

#### Exemplary Methods for Detecting Deletions and Duplications Using Allele Distributions or Patterns

**[259]** This section includes methods for determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment. In some embodiments, the method involves enumerating (i) a plurality of hypotheses specifying the number of copies of the chromosome or chromosome segment that are present in the genome of one or more cells (such as cancer cells) of the individual or (ii) a plurality of hypotheses specifying the degree of overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of one or more cells of the individual. In some embodiments, the method involves obtaining genetic data from the individual at a plurality of polymorphic loci (such as SNP loci) on the chromosome or chromosome segment. In some embodiments, a probability distribution of the expected genotypes of the individual for each of the hypotheses is created. In some embodiments, a data fit between the obtained genetic data of the individual and the probability distribution of the expected genotypes of the individual is calculated. In some embodiments, one or more hypotheses are ranked according to the data fit, and the hypothesis that is ranked the highest is selected. In some embodiments, a technique or algorithm, such as a search algorithm, is used for one or more of the following steps: calculating the data fit, ranking the hypotheses, or selecting the hypothesis that is ranked the highest. In some embodiments, the data fit is a fit to a beta-binomial distribution or a fit to a binomial distribution. In some embodiments, the technique or algorithm is selected from the group consisting of maximum likelihood estimation, maximum a-posteriori estimation, Bayesian estimation, dynamic estimation (such as dynamic Bayesian estimation), and expectation-maximization estimation. In some embodiments, the method includes applying the technique or algorithm to the obtained genetic data and the expected genetic data.

**[260]** In some embodiments, the method involves enumerating (i) a plurality of hypotheses specifying the number of copies of the chromosome or chromosome segment that are present in the genome of one or more cells (such as cancer cells) of the individual or (ii) a plurality of hypotheses

specifying the degree of overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of one or more cells of the individual. In some embodiments, the method involves obtaining genetic data from the individual at a plurality of polymorphic loci (such as SNP loci) on the chromosome or chromosome segment. In some embodiments, the genetic data includes allele counts for the plurality of polymorphic loci. In some embodiments, a joint distribution model is created for the expected allele counts at the plurality of polymorphic loci on the chromosome or chromosome segment for each hypothesis. In some embodiments, a relative probability for one or more of the hypotheses is determined using the joint distribution model and the allele counts measured on the sample, and the hypothesis with the greatest probability is selected.

**[261]** In some embodiments, the distribution or pattern of alleles (such as the pattern of calculated allele ratios) is used to determine the presence or absence of a CNV, such as a deletion or duplication. If desired the parental origin of the CNV can be determined based on this pattern. A maternally inherited duplication is an extra copy of a chromosome segment from the mother, and maternally inherited deletion is the absence of the copy of a chromosome segment from the mother such that the only copy of the chromosome segment that is present is from the father. Exemplary patterns are illustrated in FIGs. 15A-19D and are described further below.

**[262]** To determine the presence or absence of a deletion of a chromosome segment of interest, the algorithm considers the distribution of sequence counts from each of two possible alleles at large number of SNPs per chromosome. It is important to note that some embodiments of the algorithm use an approach that does not lend itself to visualization. Thus, for the purposes of illustration, the data is displayed in FIGs. 15A-18 in a simplified fashion as ratios of the two most likely alleles, labeled as A and B, so that the relevant trends can be more readily visualized. This simplified illustration does not take into account some of the possible features of the algorithm. For example, two embodiments for the algorithm that are not possible to illustrate with a method of visualization that displays allele ratios are: 1) the ability to leverage linkage disequilibrium, *i.e.* the influence that a measurement at one SNP has on the likely identity of a neighboring SNP, and 2) the use of non-Gaussian data models that describe the expected distribution of allele measurements at a SNP given platform characteristics and amplification biases. Also note that a simplified version of the algorithm only considers the two most common alleles at each SNP, ignoring other possible alleles.

**[263]** Deletions of interest were detected in genomic and maternal blood samples. In some

specifying the degree of overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of one or more cells of the individual. In some embodiments, the method involves obtaining genetic data from the individual at a plurality of polymorphic loci (such as SNP loci) on the chromosome or chromosome segment. In some embodiments, the genetic data includes allele counts for the plurality of polymorphic loci. In some embodiments, a joint distribution model is created for the expected allele counts at the plurality of polymorphic loci on the chromosome or chromosome segment for each hypothesis. In some embodiments, a relative probability for one or more of the hypotheses is determined using the joint distribution model and the allele counts measured on the sample, and the hypothesis with the greatest probability is selected.

**[261]** In some embodiments, the distribution or pattern of alleles (such as the pattern of calculated allele ratios) is used to determine the presence or absence of a CNV, such as a deletion or duplication. If desired the parental origin of the CNV can be determined based on this pattern. A maternally inherited duplication is an extra copy of a chromosome segment from the mother, and maternally inherited deletion is the absence of the copy of a chromosome segment from the mother such that the only copy of the chromosome segment that is present is from the father. Exemplary patterns are illustrated in FIGs. 15A-19D and are described further below.

**[262]** To determine the presence or absence of a deletion of a chromosome segment of interest, the algorithm considers the distribution of sequence counts from each of two possible alleles at large number of SNPs per chromosome. It is important to note that some embodiments of the algorithm use an approach that does not lend itself to visualization. Thus, for the purposes of illustration, the data is displayed in FIGs. 15A-18 in a simplified fashion as ratios of the two most likely alleles, labeled as A and B, so that the relevant trends can be more readily visualized. This simplified illustration does not take into account some of the possible features of the algorithm. For example, two embodiments for the algorithm that are not possible to illustrate with a method of visualization that displays allele ratios are: 1) the ability to leverage linkage disequilibrium, *i.e.* the influence that a measurement at one SNP has on the likely identity of a neighboring SNP, and 2) the use of non-Gaussian data models that describe the expected distribution of allele measurements at a SNP given platform characteristics and amplification biases. Also note that a simplified version of the algorithm only considers the two most common alleles at each SNP, ignoring other possible alleles.

**[263]** Deletions of interest were detected in genomic and maternal blood samples. In some

embodiments, the genomic and maternal plasma samples are analyzed using the multiplex-PCR and sequencing method of Example 1. The genomic DNA syndrome samples tested lacked heterozygous SNPs in the targeted regions, confirming the ability of the assays to distinguish monosomy (affected) from disomy (unaffected). Analysis of cfDNA from a maternal blood sample was able to detect 22q11.2 deletion syndrome, Cri-du-Chat deletion syndrome, and Wolf-Hirschhorn deletion syndrome, as well as the other deletion syndromes in FIG. 14 in the fetus.

**[264]** FIGs. 15A-15C depict data that indicate the presence of two chromosomes when the sample is entirely maternal (no fetal cfDNA present, Figure 15A), contains a moderate fetal cfDNA fraction of 12% (Figure 15B), or contains a high fetal cfDNA fraction of 26% (Figure 15C). The x-axis represents the linear position of the individual polymorphic loci along the chromosome, and the y-axis represents the number of A allele reads as a fraction of the total (A+B) allele reads. Maternal and fetal genotypes are indicated to the right of the plots. The plots are color-coded according to maternal genotype, such that red indicates a maternal genotype of

AA, blue indicates a maternal genotype of BB, and green indicates a maternal genotype of AB. Note that the measurements are made on total cfDNA isolated from maternal blood, and the cfDNA includes both maternal and fetal cfDNA; thus, each spot represents the combination of the fetal and maternal DNA contribution for that SNP. Therefore, increasing the proportion of maternal cfDNA from 0% to 100% will gradually shift some spots up or down within the plots, depending on the maternal and fetal genotype.

**[265]** In all cases, SNPs that are homozygous for the A allele (AA) in both the mother and the fetus are found tightly associated with the upper limit of the plots, as the fraction of A allele reads is high because there should be no B alleles present. Conversely, SNPs that are homozygous for the B allele in both the mother and the fetus are found tightly associated with the lower limit of the plots, as the fraction of A allele reads is low because there should be only B alleles. The spots that are not tightly associated with the upper and lower limits of the plots represent SNPs for which the mother, the fetus, or both are heterozygous; these spots are useful for identifying fetal deletions or duplications, but can also be informative for determining paternal versus maternal inheritance. These spots segregate based on both maternal and fetal genotypes and fetal fraction, and as such the precise position of each individual spot along the y-axis depends on both stoichiometry and fetal fraction. For example, loci where the mother is AA and the fetus is AB are expected to have a different fraction of A allele reads, and thus different positioning along the y-axis, depending on the fetal fraction.

**[266]** FIG. 15 A has data for a non-pregnant woman, and thus represents the pattern when the genotype

embodiments, the genomic and maternal plasma samples are analyzed using the multiplex-PCR and sequencing method of Example 1. The genomic DNA syndrome samples tested lacked heterozygous SNPs in the targeted regions, confirming the ability of the assays to distinguish monosomy (affected) from disomy (unaffected). Analysis of cfDNA from a maternal blood sample was able to detect 22q11.2 deletion syndrome, Cri-du-Chat deletion syndrome, and Wolf-Hirschhorn deletion syndrome, as well as the other deletion syndromes in FIG. 14 in the fetus.

**[264]** FIGs. 15A-15C depict data that indicate the presence of two chromosomes when the sample is entirely maternal (no fetal cfDNA present, Figure 15A), contains a moderate fetal cfDNA fraction of 12% (Figure 15B), or contains a high fetal cfDNA fraction of 26% (Figure 15C). The x-axis represents the linear position of the individual polymorphic loci along the chromosome, and the y-axis represents the number of A allele reads as a fraction of the total (A+B) allele reads. Maternal and fetal genotypes are indicated to the right of the plots. The plots are color-coded according to maternal genotype, such that red indicates a maternal genotype of

AA, blue indicates a maternal genotype of BB, and green indicates a maternal genotype of AB. Note that the measurements are made on total cfDNA isolated from maternal blood, and the cfDNA includes both maternal and fetal cfDNA; thus, each spot represents the combination of the fetal and maternal DNA contribution for that SNP. Therefore, increasing the proportion of maternal cfDNA from 0% to 100% will gradually shift some spots up or down within the plots, depending on the maternal and fetal genotype.

**[265]** In all cases, SNPs that are homozygous for the A allele (AA) in both the mother and the fetus are found tightly associated with the upper limit of the plots, as the fraction of A allele reads is high because there should be no B alleles present. Conversely, SNPs that are homozygous for the B allele in both the mother and the fetus are found tightly associated with the lower limit of the plots, as the fraction of A allele reads is low because there should be only B alleles. The spots that are not tightly associated with the upper and lower limits of the plots represent SNPs for which the mother, the fetus, or both are heterozygous; these spots are useful for identifying fetal deletions or duplications, but can also be informative for determining paternal versus maternal inheritance. These spots segregate based on both maternal and fetal genotypes and fetal fraction, and as such the precise position of each individual spot along the y-axis depends on both stoichiometry and fetal fraction. For example, loci where the mother is AA and the fetus is AB are expected to have a different fraction of A allele reads, and thus different positioning along the y-axis, depending on the fetal fraction.

**[266]** FIG. 15 A has data for a non-pregnant woman, and thus represents the pattern when the genotype

is entirely maternal. This pattern includes “clusters” of spots: a red cluster tightly associated with the top of the plot (SNPs where the maternal genotype is AA), a blue cluster tightly associated with the bottom of the plot (SNPs where the maternal genotype is BB), and a single, centered green cluster (SNPs where the maternal genotype is AB). For FIG. 15B, the contribution of fetal alleles to the fraction of A allele reads shifts the position of some allele spots up or down along the y-axis. For FIG. 15C, the pattern, including two red and two blue peripheral bands and a trio of central green bands, is readily apparent. The three central green bands correspond to SNPs that are heterozygous in the mother, and two “peripheral” bands each at both the top (red) and bottom (blue) of the plots correspond to SNPs that are homozygous in the mother.

**[267]** Analysis of a 22q11.2 deletion carrier (a mother with this deletion) is shown in FIG. 16A. The deletion carrier does not have heterozygous SNPs in this region since the carrier only has one copy of this region. Thus, this deletion is indicated by the absence of the green AB SNPs. The analysis of a paternally inherited 22q11 deletion in a fetus is shown in FIG. 16B. When the fetus only inherits a single copy of a chromosome segment (in the case of a paternally inherited deletion, the copy present in the fetus comes from the mother), and thus only inherits a single allele for each locus in this segment, heterozygosity of the fetus is not possible. As such, the only possible fetal SNP identities are A or B. Note the absence of internal peripheral bands. For a paternally inherited deletion, the characteristic pattern includes two central green bands that represent SNPs for which the mother is heterozygous, and only has single peripheral red and blue bands that represent SNPs for which the mother is homozygous, and which remain tightly associated with the upper and lower limits of the plots (1 and 0), respectively.

**[268]** Analysis of a maternally inherited Cri-du-Chat deletion syndrome is shown in FIG. 17. There are two central green bands instead of three green bands, and there are two red and two blue peripheral bands. A maternally inherited deletion (such as a maternal carrier of Duchenne's muscular dystrophy) can also be detected based on the small amount of signal in that region of the deletion in a mixed sample of maternal and fetal DNA (such as a plasma sample) due to both the mother and the fetus having the deletion.

**[269]** FIG. 18 is a plot of a paternally inherited Wolf-Hirschhorn deletion syndrome, as indicated by the presence of one red and one blue peripheral band.

**[270]** If desired, similar plots can be generated for a sample from an individual suspected of having a deletion or duplication, such as a CNV associated with cancer. In such plots, the following color coding can be used based on the genotype of cells without the CNV: red indicates a genotype of AA, blue

is entirely maternal. This pattern includes “clusters” of spots: a red cluster tightly associated with the top of the plot (SNPs where the maternal genotype is AA), a blue cluster tightly associated with the bottom of the plot (SNPs where the maternal genotype is BB), and a single, centered green cluster (SNPs where the maternal genotype is AB). For FIG. 15B, the contribution of fetal alleles to the fraction of A allele reads shifts the position of some allele spots up or down along the y-axis. For FIG. 15C, the pattern, including two red and two blue peripheral bands and a trio of central green bands, is readily apparent. The three central green bands correspond to SNPs that are heterozygous in the mother, and two “peripheral” bands each at both the top (red) and bottom (blue) of the plots correspond to SNPs that are homozygous in the mother.

**[267]** Analysis of a 22q11.2 deletion carrier (a mother with this deletion) is shown in FIG. 16A. The deletion carrier does not have heterozygous SNPs in this region since the carrier only has one copy of this region. Thus, this deletion is indicated by the absence of the green AB SNPs. The analysis of a paternally inherited 22q11 deletion in a fetus is shown in FIG. 16B. When the fetus only inherits a single copy of a chromosome segment (in the case of a paternally inherited deletion, the copy present in the fetus comes from the mother), and thus only inherits a single allele for each locus in this segment, heterozygosity of the fetus is not possible. As such, the only possible fetal SNP identities are A or B. Note the absence of internal peripheral bands. For a paternally inherited deletion, the characteristic pattern includes two central green bands that represent SNPs for which the mother is heterozygous, and only has single peripheral red and blue bands that represent SNPs for which the mother is homozygous, and which remain tightly associated with the upper and lower limits of the plots (1 and 0), respectively.

**[268]** Analysis of a maternally inherited Cri-du-Chat deletion syndrome is shown in FIG. 17. There are two central green bands instead of three green bands, and there are two red and two blue peripheral bands. A maternally inherited deletion (such as a maternal carrier of Duchenne's muscular dystrophy) can also be detected based on the small amount of signal in that region of the deletion in a mixed sample of maternal and fetal DNA (such as a plasma sample) due to both the mother and the fetus having the deletion.

**[269]** FIG. 18 is a plot of a paternally inherited Wolf-Hirschhorn deletion syndrome, as indicated by the presence of one red and one blue peripheral band.

**[270]** If desired, similar plots can be generated for a sample from an individual suspected of having a deletion or duplication, such as a CNV associated with cancer. In such plots, the following color coding can be used based on the genotype of cells without the CNV: red indicates a genotype of AA, blue

indicates a genotype of BB, and green indicates a genotype of

AB. In some embodiments for a deletion, the pattern includes two central green bands that represent SNPs for which the individual is heterozygous (top green band represents AB from cells without the deletion and A from cells with the deletion, and bottom green band represents AB from cells without the deletion and B from cells with the deletion), and only has single peripheral red and blue bands that represent SNPs for which the individual is homozygous, and which remain tightly associated with the upper and lower limits of the plots (1 and 0), respectively. In some embodiments, the separation of the two green bands increases as the fraction of cells, DNA, or RNA with the deletion increases.

#### Exemplary Methods for Identifying and Analyzing Multiple Pregnancies

[271] In some embodiments, any of the methods of the present invention are used to detect the presence of a multiple pregnancy, such as a twin pregnancy, where at least one of the fetuses is genetically different from at least one other fetus. In some embodiments, fraternal twins are identified based on the presence of two fetus with different allele, different allele ratios, or different allele distributions at some (or all) of the tested loci. In some embodiments, fraternal twins are identified by determining the expected allele ratio at each locus (such as SNP loci) for two fetuses that may have the same or different fetal fractions in the sample (such as a plasma sample). In some embodiments, the likelihood of a particular pair of fetal fractions (where  $f_1$  is the fetal fraction for fetus 1, and  $f_2$  is the fetal fraction for fetus 2) is calculated by considering some or all of the possible genotypes of the two fetuses, conditioned on the mother's genotype and genotype population frequencies. The mixture of two fetal and one maternal genotype, combined with the fetal fractions, determine the expected allele ratio at a SNP. For example, if the mother is AA, fetus 1 is AA, and fetus 2 is AB, the overall fraction of B allele at the SNP is one-half of  $f_2$ . The likelihood calculation asks how well all of the SNPs together match the expected allele ratios based on all of the possible combinations of fetal genotypes. The fetal fraction pair ( $f_1$ ,  $f_2$ ) that best matches the data is selected. It is not necessary to calculate specific genotypes of the fetuses; instead, one can, for example, consider all of the possible genotypes in a statistical combination. In some embodiments, if the method does not distinguish between singleton and identical twins, an ultrasound can be performed to determine whether there is a singleton or identical twin pregnancy. If the ultrasound detects a twin pregnancy it can be assumed that the pregnancy is an identical twin pregnancy because a fraternal twin pregnancy would have been detected based on the SNP analysis discussed above.

[272] In some embodiments, a pregnant mother is known to have a multiple pregnancy (such as a twin pregnancy) based on prior testing, such as an ultrasound. Any of the methods of the present invention can

indicates a genotype of BB, and green indicates a genotype of

AB. In some embodiments for a deletion, the pattern includes two central green bands that represent SNPs for which the individual is heterozygous (top green band represents AB from cells without the deletion and A from cells with the deletion, and bottom green band represents AB from cells without the deletion and B from cells with the deletion), and only has single peripheral red and blue bands that represent SNPs for which the individual is homozygous, and which remain tightly associated with the upper and lower limits of the plots (1 and 0), respectively. In some embodiments, the separation of the two green bands increases as the fraction of cells, DNA, or RNA with the deletion increases.

#### Exemplary Methods for Identifying and Analyzing Multiple Pregnancies

[271] In some embodiments, any of the methods of the present invention are used to detect the presence of a multiple pregnancy, such as a twin pregnancy, where at least one of the fetuses is genetically different from at least one other fetus. In some embodiments, fraternal twins are identified based on the presence of two fetus with different allele, different allele ratios, or different allele distributions at some (or all) of the tested loci. In some embodiments, fraternal twins are identified by determining the expected allele ratio at each locus (such as SNP loci) for two fetuses that may have the same or different fetal fractions in the sample (such as a plasma sample). In some embodiments, the likelihood of a particular pair of fetal fractions (where  $f_1$  is the fetal fraction for fetus 1, and  $f_2$  is the fetal fraction for fetus 2) is calculated by considering some or all of the possible genotypes of the two fetuses, conditioned on the mother's genotype and genotype population frequencies. The mixture of two fetal and one maternal genotype, combined with the fetal fractions, determine the expected allele ratio at a SNP. For example, if the mother is AA, fetus 1 is AA, and fetus 2 is AB, the overall fraction of B allele at the SNP is one-half of  $f_2$ . The likelihood calculation asks how well all of the SNPs together match the expected allele ratios based on all of the possible combinations of fetal genotypes. The fetal fraction pair ( $f_1$ ,  $f_2$ ) that best matches the data is selected. It is not necessary to calculate specific genotypes of the fetuses; instead, one can, for example, consider all of the possible genotypes in a statistical combination. In some embodiments, if the method does not distinguish between singleton and identical twins, an ultrasound can be performed to determine whether there is a singleton or identical twin pregnancy. If the ultrasound detects a twin pregnancy it can be assumed that the pregnancy is an identical twin pregnancy because a fraternal twin pregnancy would have been detected based on the SNP analysis discussed above.

[272] In some embodiments, a pregnant mother is known to have a multiple pregnancy (such as a twin pregnancy) based on prior testing, such as an ultrasound. Any of the methods of the present invention can

be used to determine whether the multiple pregnancy includes identical or fraternal twins. For example, the measured allele ratios can be compared to what would be expected for identical twins (the same allele ratios as a singleton pregnancy) or for fraternal twins (such as the calculation of allele ratios as described above). Some identical twins are monochorionic twins, which have a risk of twin-to-twin transfusion syndrome. Thus, twins determined to be identical twins using a method of the invention are desirably tested (such as by ultrasound) to determine if they are monochorionic twins, and if so, these twins can be monitored (such as bi-weekly ultrasounds from 16 weeks) for signs of win-to-twin transfusion syndrome.

**[273]** In some embodiments, any of the methods of the present invention are used to determine whether any of the fetuses in a multiple pregnancy, such as a twin pregnancy, are aneuploid. Aneuploidy testing for twins begins with the fetal fraction estimate. In some embodiments, the fetal fraction pair ( $f_1$ ,  $f_2$ ) that best matches the data is selected as described above. In some embodiments, a maximum likelihood estimate is performed for the parameter pair ( $f_1$ ,  $f_2$ ) over the range of possible fetal fractions. In some embodiments, the range of  $f_2$  is from 0 to  $f_1$  because  $f_2$  is defined as the smaller fetal fraction. Given a pair ( $f_1$ ,  $f_2$ ), data likelihood is calculated from the allele ratios observed at a set of loci such as SNP loci. In some embodiments, the data likelihood reflects the genotypes of the mother, the father if available, population frequencies, and the resulting probabilities of fetal genotypes. In some embodiments, SNPs are assumed independent. The estimated fetal fraction pair is the one that produces the highest data likelihood. If  $f_2$  is 0 then the data is best explained by only one set of fetal genotypes, indicating identical twins, where  $f_1$  is the combined fetal fraction. Otherwise  $f_1$  and  $f_2$  are the estimates of the individual twin fetal fractions. Having established the best estimate of ( $f_1$ ,  $f_2$ ), one can predict the overall fraction of B allele in the plasma for any combination of maternal and fetal genotypes, if desired. It is not necessary to assign individual sequence reads to the individual fetuses. Ploidy testing is performed using another maximum likelihood estimate which compares the data likelihood of two hypotheses. In some embodiments for identical twins, one consider the hypotheses (i) both twins are euploid, and (ii) both twins are trisomic. In some embodiments for fraternal twins, one considers the hypotheses (i) both twins are euploid and (ii) at least one twin is trisomic. The trisomy hypotheses for fraternal twins are based on the lower fetal fraction, since a trisomy in the twin with a higher fetal fraction would also be detected. Ploidy likelihoods are calculated using a method which predicts the expected number of reads at each targeted genome locus conditioned on either the disomy or trisomy hypothesis. There is no requirement for a disomy reference chromosome. The variance model for the expected number of reads takes into account the performance of individual target loci as well as the correlation between loci (see, for example,

be used to determine whether the multiple pregnancy includes identical or fraternal twins. For example, the measured allele ratios can be compared to what would be expected for identical twins (the same allele ratios as a singleton pregnancy) or for fraternal twins (such as the calculation of allele ratios as described above). Some identical twins are monochorionic twins, which have a risk of twin-to-twin transfusion syndrome. Thus, twins determined to be identical twins using a method of the invention are desirably tested (such as by ultrasound) to determine if they are monochorionic twins, and if so, these twins can be monitored (such as bi-weekly ultrasounds from 16 weeks) for signs of win-to-twin transfusion syndrome.

**[273]** In some embodiments, any of the methods of the present invention are used to determine whether any of the fetuses in a multiple pregnancy, such as a twin pregnancy, are aneuploid. Aneuploidy testing for twins begins with the fetal fraction estimate. In some embodiments, the fetal fraction pair ( $f_1$ ,  $f_2$ ) that best matches the data is selected as described above. In some embodiments, a maximum likelihood estimate is performed for the parameter pair ( $f_1$ ,  $f_2$ ) over the range of possible fetal fractions. In some embodiments, the range of  $f_2$  is from 0 to  $f_1$  because  $f_2$  is defined as the smaller fetal fraction. Given a pair ( $f_1$ ,  $f_2$ ), data likelihood is calculated from the allele ratios observed at a set of loci such as SNP loci. In some embodiments, the data likelihood reflects the genotypes of the mother, the father if available, population frequencies, and the resulting probabilities of fetal genotypes. In some embodiments, SNPs are assumed independent. The estimated fetal fraction pair is the one that produces the highest data likelihood. If  $f_2$  is 0 then the data is best explained by only one set of fetal genotypes, indicating identical twins, where  $f_1$  is the combined fetal fraction. Otherwise  $f_1$  and  $f_2$  are the estimates of the individual twin fetal fractions. Having established the best estimate of ( $f_1$ ,  $f_2$ ), one can predict the overall fraction of B allele in the plasma for any combination of maternal and fetal genotypes, if desired. It is not necessary to assign individual sequence reads to the individual fetuses. Ploidy testing is performed using another maximum likelihood estimate which compares the data likelihood of two hypotheses. In some embodiments for identical twins, one consider the hypotheses (i) both twins are euploid, and (ii) both twins are trisomic. In some embodiments for fraternal twins, one considers the hypotheses (i) both twins are euploid and (ii) at least one twin is trisomic. The trisomy hypotheses for fraternal twins are based on the lower fetal fraction, since a trisomy in the twin with a higher fetal fraction would also be detected. Ploidy likelihoods are calculated using a method which predicts the expected number of reads at each targeted genome locus conditioned on either the disomy or trisomy hypothesis. There is no requirement for a disomy reference chromosome. The variance model for the expected number of reads takes into account the performance of individual target loci as well as the correlation between loci (see, for example,

U.S. Serial No. 62/008,235, filed June 5, 2014, and U.S. Serial No. 62/032,785, filed August 4, 2014, which are each hereby incorporated by reference in its entirety). If the smaller twin has fetal fraction  $f_1$ , our ability to detect a trisomy in that twin is equivalent to our ability to detect a trisomy in a singleton pregnancy at the same fetal fraction. This is because the part of the method that detects the trisomy in some embodiments does not depend on genotypes and does not distinguish between multiple or singleton pregnancy. It simply looks for an increased number of reads in accordance with the determined fetal fraction.

**[274]** In some embodiments, the method includes detecting the presence of twins based on SNP loci (such as described above). If twins are detected, SPNs are used to determine the fetal fraction of each fetus ( $f_1$ ,  $f_2$ ) such as described above. In some embodiments, samples that have high confidence disomy calls are used to determine the amplification bias on a per-SNP basis. In some embodiments, these samples with high confidence disomy calls are analyzed in the same run as one or more samples of interest. In some embodiments, the amplification bias on a per-SNP basis is used to model the distribution of reads for one or more chromosomes or chromosome segments of interest such as chromosome 21 that are expected or the disomy hypothesis and the trisomy hypothesis given the lower of the two twin fetal fraction. The likelihood or probability of disomy or trisomy is calculated given the two models and the measured quantity of the chromosome or chromosome segment of interest.

**[275]** In some embodiments, the threshold for a positive aneuploidy call (such as a trisomy call) is set based on the twin with the lower fetal fraction. This way, if the other twin is positive, or if both are positive, the total chromosome representation is definitely above the threshold.

#### Exemplary Counting Methods/Quantitative Methods

**[276]** In some embodiments, one or more counting methods (also referred to as quantitative methods) are used to detect one or more CNS, such as deletions or duplications of chromosome segments or entire chromosomes. In some embodiments, one or more counting methods are used to determine whether the overrepresentation of the number of copies of the first homologous chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment. In some embodiments, one or more counting methods are used to determine the number of extra copies of a chromosome segment or chromosome that is duplicated (such as whether there are 1, 2, 3, 4, or more extra copies). In some embodiments, one or more counting methods are used to differentiate a sample has many duplications and a smaller tumor fraction from a sample with fewer duplications and a larger tumor fraction. For example, one or more counting methods may be used to

U.S. Serial No. 62/008,235, filed June 5, 2014, and U.S. Serial No. 62/032,785, filed August 4, 2014, which are each hereby incorporated by reference in its entirety). If the smaller twin has fetal fraction  $f_1$ , our ability to detect a trisomy in that twin is equivalent to our ability to detect a trisomy in a singleton pregnancy at the same fetal fraction. This is because the part of the method that detects the trisomy in some embodiments does not depend on genotypes and does not distinguish between multiple or singleton pregnancy. It simply looks for an increased number of reads in accordance with the determined fetal fraction.

**[274]** In some embodiments, the method includes detecting the presence of twins based on SNP loci (such as described above). If twins are detected, SPNs are used to determine the fetal fraction of each fetus ( $f_1$ ,  $f_2$ ) such as described above. In some embodiments, samples that have high confidence disomy calls are used to determine the amplification bias on a per-SNP basis. In some embodiments, these samples with high confidence disomy calls are analyzed in the same run as one or more samples of interest. In some embodiments, the amplification bias on a per-SNP basis is used to model the distribution of reads for one or more chromosomes or chromosome segments of interest such as chromosome 21 that are expected or the disomy hypothesis and the trisomy hypothesis given the lower of the two twin fetal fraction. The likelihood or probability of disomy or trisomy is calculated given the two models and the measured quantity of the chromosome or chromosome segment of interest.

**[275]** In some embodiments, the threshold for a positive aneuploidy call (such as a trisomy call) is set based on the twin with the lower fetal fraction. This way, if the other twin is positive, or if both are positive, the total chromosome representation is definitely above the threshold.

#### Exemplary Counting Methods/Quantitative Methods

**[276]** In some embodiments, one or more counting methods (also referred to as quantitative methods) are used to detect one or more CNS, such as deletions or duplications of chromosome segments or entire chromosomes. In some embodiments, one or more counting methods are used to determine whether the overrepresentation of the number of copies of the first homologous chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment. In some embodiments, one or more counting methods are used to determine the number of extra copies of a chromosome segment or chromosome that is duplicated (such as whether there are 1, 2, 3, 4, or more extra copies). In some embodiments, one or more counting methods are used to differentiate a sample has many duplications and a smaller tumor fraction from a sample with fewer duplications and a larger tumor fraction. For example, one or more counting methods may be used to

differentiate a sample with four extra chromosome copies and a tumor fraction of 10% from a sample with two extra chromosome copies and a tumor fraction of 20%. Exemplary methods are disclosed, *e.g.* U.S. Publication Nos. 2007/0184467; 2013/0172211; and 2012/0003637; U.S. Patent Nos. 8,467,976; 7,888,017; 8,008,018; 8,296,076; and 8,195,415; U.S. Serial No. 62/008,235, filed June 5, 2014, and U.S. Serial No. 62/032,785, filed August 4, 2014, which are each hereby incorporated by reference in its entirety.

**[277]** In some embodiment, the counting method includes counting the number of DNA sequence-based reads that map to one or more given chromosomes or chromosome segments. Some such methods involve creation of a reference value (cut-off value) for the number of DNA sequence reads mapping to a specific chromosome or chromosome segment, wherein a number of reads in excess of the value is indicative of a specific genetic abnormality.

**[278]** In some embodiments, the total measured quantity of all the alleles for one or more loci (such as the total amount of a polymorphic or non-polymorphic locus) is compared to a reference amount. In some embodiments, the reference amount is (i) a threshold value or (ii) an expected amount for a particular copy number hypothesis. In some embodiments, the reference amount (for the absence of a CNV) is the total measured quantity of all the alleles for one or more loci for one or more chromosomes or chromosomes segments known or expected to not have a deletion or duplication. In some embodiments, the reference amount (for the presence of a CNV) is the total measured quantity of all the alleles for one or more loci for one or more chromosomes or chromosomes segments known or expected to have a deletion or duplication. In some embodiments, the reference amount is the total measured quantity of all the alleles for one or more loci for one or more reference chromosomes or chromosome segments. In some embodiments, the reference amount is the mean or median of the values determined for two or more different chromosomes, chromosome segments, or different samples. In some embodiments, random (*e.g.*, massively parallel shotgun sequencing) or targeted sequencing is used to determine the amount of one or more polymorphic or non-polymorphic loci.

**[279]** In some embodiments utilizing a reference amount, the method includes (a) measuring the amount of genetic material on a chromosome or chromosome segment of interest; (b) comparing the amount from step (a) to a reference amount; and (c) identifying the presence or absence of a deletion or duplication based on the comparison.

**[280]** In some embodiments utilizing a reference chromosome or chromosome segment, the method includes sequencing DNA or RNA from a sample to obtain a plurality of sequence tags aligning to target

differentiate a sample with four extra chromosome copies and a tumor fraction of 10% from a sample with two extra chromosome copies and a tumor fraction of 20%. Exemplary methods are disclosed, *e.g.* U.S. Publication Nos. 2007/0184467; 2013/0172211; and 2012/0003637; U.S. Patent Nos. 8,467,976; 7,888,017; 8,008,018; 8,296,076; and 8,195,415; U.S. Serial No. 62/008,235, filed June 5, 2014, and U.S. Serial No. 62/032,785, filed August 4, 2014, which are each hereby incorporated by reference in its entirety.

**[277]** In some embodiment, the counting method includes counting the number of DNA sequence-based reads that map to one or more given chromosomes or chromosome segments. Some such methods involve creation of a reference value (cut-off value) for the number of DNA sequence reads mapping to a specific chromosome or chromosome segment, wherein a number of reads in excess of the value is indicative of a specific genetic abnormality.

**[278]** In some embodiments, the total measured quantity of all the alleles for one or more loci (such as the total amount of a polymorphic or non-polymorphic locus) is compared to a reference amount. In some embodiments, the reference amount is (i) a threshold value or (ii) an expected amount for a particular copy number hypothesis. In some embodiments, the reference amount (for the absence of a CNV) is the total measured quantity of all the alleles for one or more loci for one or more chromosomes or chromosomes segments known or expected to not have a deletion or duplication. In some embodiments, the reference amount (for the presence of a CNV) is the total measured quantity of all the alleles for one or more loci for one or more chromosomes or chromosomes segments known or expected to have a deletion or duplication. In some embodiments, the reference amount is the total measured quantity of all the alleles for one or more loci for one or more reference chromosomes or chromosome segments. In some embodiments, the reference amount is the mean or median of the values determined for two or more different chromosomes, chromosome segments, or different samples. In some embodiments, random (*e.g.*, massively parallel shotgun sequencing) or targeted sequencing is used to determine the amount of one or more polymorphic or non-polymorphic loci.

**[279]** In some embodiments utilizing a reference amount, the method includes (a) measuring the amount of genetic material on a chromosome or chromosome segment of interest; (b) comparing the amount from step (a) to a reference amount; and (c) identifying the presence or absence of a deletion or duplication based on the comparison.

**[280]** In some embodiments utilizing a reference chromosome or chromosome segment, the method includes sequencing DNA or RNA from a sample to obtain a plurality of sequence tags aligning to target

loci. In some embodiments, the sequence tags are of sufficient length to be assigned to a specific target locus (*e.g.*, 15-100 nucleotides in length); the target loci are from a plurality of different chromosomes or chromosome segments that include at least one first chromosome or chromosome segment suspected of having an abnormal distribution in the sample and at least one second chromosome or chromosome segment presumed to be normally distributed in the sample. In some embodiments, the plurality of sequence tags are assigned to their corresponding target loci. In some embodiments, the number of sequence tags aligning to the target loci of the first chromosome or chromosome segment and the number of sequence tags aligning to the target loci of the second chromosome or chromosome segment are determined. In some embodiments, these numbers are compared to determine the presence or absence of an abnormal distribution (such as a deletion or duplication) of the first chromosome or chromosome segment.

**[281]** In some embodiments, the value of/(such as the fetal fraction or tumor fraction) is used in the CNV determination, such as to compare the observed difference between the amount of two chromosomes or chromosome segments to the difference that would be expected for a particular type of CNV given the value of/(see, *e.g.*, US Publication No 2012/0190020; US Publication No 2012/0190021; US Publication No 2012/0190557; US Publication No 2012/0191358, which are each hereby incorporated by reference in its entirety). For example, the difference in the amount of a chromosome segment that is duplicated in a fetus compared to a disomic reference chromosome segment in a blood sample from a mother carrying the fetus increases as the fetal fraction increases. Additionally, the difference in the amount of a chromosome segment that is duplicated in a tumor compared to a disomic reference chromosome segment increases as the tumor fraction increases. In some embodiments, the method includes comparing the relative frequency of a chromosome or chromosome segment of interest to a reference chromosomes or chromosome segment (such as a chromosome or chromosome segment expected or known to be disomic) to the value of/ to determine the likelihood of the CNV. For example, the difference in amounts between the first chromosomes or chromosome segment to the reference chromosome or chromosome segment can be compared to what would be expected given the value of/ for various possible CNVs (such as one or two extra copies of a chromosome segment of interest).

**[282]** The following prophetic examples illustrate the use of a counting method/quantitative method to differentiate between a duplication of the first homologous chromosome segment and a deletion of the second homologous chromosome segment. If one considers the normal disomic genome of the host to be the baseline, then analysis of a mixture of normal and cancer cells yields the average difference between

loci. In some embodiments, the sequence tags are of sufficient length to be assigned to a specific target locus (*e.g.*, 15-100 nucleotides in length); the target loci are from a plurality of different chromosomes or chromosome segments that include at least one first chromosome or chromosome segment suspected of having an abnormal distribution in the sample and at least one second chromosome or chromosome segment presumed to be normally distributed in the sample. In some embodiments, the plurality of sequence tags are assigned to their corresponding target loci. In some embodiments, the number of sequence tags aligning to the target loci of the first chromosome or chromosome segment and the number of sequence tags aligning to the target loci of the second chromosome or chromosome segment are determined. In some embodiments, these numbers are compared to determine the presence or absence of an abnormal distribution (such as a deletion or duplication) of the first chromosome or chromosome segment.

**[281]** In some embodiments, the value of/(such as the fetal fraction or tumor fraction) is used in the CNV determination, such as to compare the observed difference between the amount of two chromosomes or chromosome segments to the difference that would be expected for a particular type of CNV given the value of/(see, *e.g.*, US Publication No 2012/0190020; US Publication No 2012/0190021; US Publication No 2012/0190557; US Publication No 2012/0191358, which are each hereby incorporated by reference in its entirety). For example, the difference in the amount of a chromosome segment that is duplicated in a fetus compared to a disomic reference chromosome segment in a blood sample from a mother carrying the fetus increases as the fetal fraction increases. Additionally, the difference in the amount of a chromosome segment that is duplicated in a tumor compared to a disomic reference chromosome segment increases as the tumor fraction increases. In some embodiments, the method includes comparing the relative frequency of a chromosome or chromosome segment of interest to a reference chromosomes or chromosome segment (such as a chromosome or chromosome segment expected or known to be disomic) to the value of/ to determine the likelihood of the CNV. For example, the difference in amounts between the first chromosomes or chromosome segment to the reference chromosome or chromosome segment can be compared to what would be expected given the value of/ for various possible CNVs (such as one or two extra copies of a chromosome segment of interest).

**[282]** The following prophetic examples illustrate the use of a counting method/quantitative method to differentiate between a duplication of the first homologous chromosome segment and a deletion of the second homologous chromosome segment. If one considers the normal disomic genome of the host to be the baseline, then analysis of a mixture of normal and cancer cells yields the average difference between

the baseline and the cancer DNA in the mixture. For example, imagine a case where 10% of the DNA in the sample originated from cells with a deletion over a region of a chromosome that is targeted by the assay. In some embodiments, a quantitative approach shows that the quantity of reads corresponding to that region is expected to be 95% of what is expected for a normal sample. This is because one of the two target chromosomal regions in each of the tumor cells with a deletion of the targeted region is missing, and thus the total amount of DNA mapping to that region is 90% (for the normal cells) plus  $10\%$  (for the tumor cells) = 95%. Alternately in some embodiments, an allelic approach shows that the ratio of alleles at heterozygous loci averaged 19:20. Now imagine a case where 10% of the DNA in the sample originated from cells with a five-fold focal amplification of a region of a chromosome that is targeted by the assay. In some embodiments, a quantitative approach shows that the quantity of reads corresponding to that region is expected to be 125% of what is expected for a normal sample. This is because one of the two target chromosomal regions in each of the tumor cells with a five-fold focal amplification is copied an extra five times over the targeted region, and thus the total amount of DNA mapping to that region is 90% (for the normal cells) plus  $(2 + 5) \times 10\% / 2$  (for the tumor cells) = 125%. Alternately in some embodiments, an allelic approach shows that the ratio of alleles at heterozygous loci averaged 25:20. Note that when using an allelic approach alone, a focal amplification of five-fold over a chromosomal region in a sample with 10% cfDNA may appear the same as a deletion over the same region in a sample with 40% cfDNA; in these two cases, the haplotype that is under-represented in the case of the deletion appears to be the haplotype without a CNV in the case with the focal duplication, and the haplotype without a CNV in the case of the deletion appears to be the over-represented haplotype in the case with the focal duplication. Combining the likelihoods produced by this allelic approach with likelihoods produced by a quantitative approach differentiates between the two possibilities.

#### Exemplary Counting Methods/Quantitative Methods Using Reference Samples

**[283]** An exemplary quantitative method that uses one or more reference samples is described in U.S. Serial No. 62/008,235, filed June 5, 2014 and U.S. Serial No. 62/032,785, filed August 4, 2014, which is hereby incorporated by reference in its entirety. In some embodiments, one or more reference samples most likely to not have any CNVs on one or more chromosomes or chromosomes of interest (*e.g.*, a normal sample) are identified by selecting the samples with the highest fraction of tumor DNA, selecting the samples with the z-score closest to zero, selecting the samples where the data fits the hypothesis corresponding to no CNVs with the highest confidence or likelihood, selecting the samples known to be normal, selecting the samples from individuals with the lowest likelihood of having cancer (*e.g.*, having

the baseline and the cancer DNA in the mixture. For example, imagine a case where 10% of the DNA in the sample originated from cells with a deletion over a region of a chromosome that is targeted by the assay. In some embodiments, a quantitative approach shows that the quantity of reads corresponding to that region is expected to be 95% of what is expected for a normal sample. This is because one of the two target chromosomal regions in each of the tumor cells with a deletion of the targeted region is missing, and thus the total amount of DNA mapping to that region is 90% (for the normal cells) plus  $10\% \times 10\%$  (for the tumor cells) = 95%. Alternately in some embodiments, an allelic approach shows that the ratio of alleles at heterozygous loci averaged 19:20. Now imagine a case where 10% of the DNA in the sample originated from cells with a five-fold focal amplification of a region of a chromosome that is targeted by the assay. In some embodiments, a quantitative approach shows that the quantity of reads corresponding to that region is expected to be 125% of what is expected for a normal sample. This is because one of the two target chromosomal regions in each of the tumor cells with a five-fold focal amplification is copied an extra five times over the targeted region, and thus the total amount of DNA mapping to that region is 90% (for the normal cells) plus  $(2 + 5) \times 10\% / 2$  (for the tumor cells) = 125%. Alternately in some embodiments, an allelic approach shows that the ratio of alleles at heterozygous loci averaged 25:20. Note that when using an allelic approach alone, a focal amplification of five-fold over a chromosomal region in a sample with 10% cfDNA may appear the same as a deletion over the same region in a sample with 40% cfDNA; in these two cases, the haplotype that is under-represented in the case of the deletion appears to be the haplotype without a CNV in the case with the focal duplication, and the haplotype without a CNV in the case of the deletion appears to be the over-represented haplotype in the case with the focal duplication. Combining the likelihoods produced by this allelic approach with likelihoods produced by a quantitative approach differentiates between the two possibilities.

#### Exemplary Counting Methods/Quantitative Methods Using Reference Samples

**[283]** An exemplary quantitative method that uses one or more reference samples is described in U.S. Serial No. 62/008,235, filed June 5, 2014 and U.S. Serial No. 62/032,785, filed August 4, 2014, which is hereby incorporated by reference in its entirety. In some embodiments, one or more reference samples most likely to not have any CNVs on one or more chromosomes or chromosomes of interest (*e.g.*, a normal sample) are identified by selecting the samples with the highest fraction of tumor DNA, selecting the samples with the z-score closest to zero, selecting the samples where the data fits the hypothesis corresponding to no CNVs with the highest confidence or likelihood, selecting the samples known to be normal, selecting the samples from individuals with the lowest likelihood of having cancer (*e.g.*, having

a low age, being a male when screening for breast cancer, having no family history, *etc.*), selecting the samples with the highest input amount of DNA, selecting the samples with the highest signal to noise ratio, selecting samples based on other criteria believed to be correlated to the likelihood of having cancer, or selecting samples using some combination of criteria. Once the reference set is chosen, one can make the assumption that these cases are disomic, and then estimate the per- SNP bias, that is, the experiment-specific amplification and other processing bias for each locus. Then, one can use this experiment-specific bias estimate to correct the bias in the measurements of the chromosome of interest, such as chromosome 21 loci, and for the other chromosome loci as appropriate, for the samples that are not part of the subset where disomy is assumed for chromosome 21. Once the biases have been corrected for in these samples of unknown ploidy, the data for these samples can then be analyzed a second time using the same or a different method to determine whether the individuals (such as fetuses) are afflicted with trisomy 21. For example, a quantitative method can be used on the remaining samples of unknown ploidy, and a z-score can be calculated using the corrected measured genetic data on chromosome 21. Alternately, as part of the preliminary estimate of the ploidy state of chromosome 21, a fetal fraction (or tumor fraction for samples from an individual suspected of having cancer) can be calculated. The proportion of corrected reads that are expected in the case of a disomy (the disomy hypothesis), and the proportion of corrected reads that are expected in the case of a trisomy (the trisomy hypothesis) can be calculated for a case with that fetal fraction. Alternately, if the fetal fraction was not measured previously, a set of disomy and trisomy hypotheses can be generated for different fetal fractions. For each case, an expected distribution of the proportion of corrected reads can be calculated given expected statistical variation in the selection and measurement of the various DNA loci. The observed corrected proportion of reads can be compared to the distribution of the expected proportion of corrected reads, and a likelihood ratio can be calculated for the disomy and trisomy hypotheses, for each of the samples of unknown ploidy. The ploidy state associated with the hypothesis with the highest calculated likelihood can be selected as the correct ploidy state.

**[284]** In some embodiments, a subset of the samples with a sufficiently low likelihood of having cancer may be selected to act as a control set of samples. The subset can be a fixed number, or it can be a variable number that is based on choosing only those samples that fall below a threshold. The quantitative data from the subset of samples may be combined, averaged, or combined using a weighted average where the weighting is based on the likelihood of the sample being normal. The quantitative data may be used to determine the per-locus bias for the amplification the sequencing of samples in the instant batch of

a low age, being a male when screening for breast cancer, having no family history, *etc.*), selecting the samples with the highest input amount of DNA, selecting the samples with the highest signal to noise ratio, selecting samples based on other criteria believed to be correlated to the likelihood of having cancer, or selecting samples using some combination of criteria. Once the reference set is chosen, one can make the assumption that these cases are disomic, and then estimate the per- SNP bias, that is, the experiment-specific amplification and other processing bias for each locus. Then, one can use this experiment-specific bias estimate to correct the bias in the measurements of the chromosome of interest, such as chromosome 21 loci, and for the other chromosome loci as appropriate, for the samples that are not part of the subset where disomy is assumed for chromosome 21. Once the biases have been corrected for in these samples of unknown ploidy, the data for these samples can then be analyzed a second time using the same or a different method to determine whether the individuals (such as fetuses) are afflicted with trisomy 21. For example, a quantitative method can be used on the remaining samples of unknown ploidy, and a z-score can be calculated using the corrected measured genetic data on chromosome 21. Alternately, as part of the preliminary estimate of the ploidy state of chromosome 21, a fetal fraction (or tumor fraction for samples from an individual suspected of having cancer) can be calculated. The proportion of corrected reads that are expected in the case of a disomy (the disomy hypothesis), and the proportion of corrected reads that are expected in the case of a trisomy (the trisomy hypothesis) can be calculated for a case with that fetal fraction. Alternately, if the fetal fraction was not measured previously, a set of disomy and trisomy hypotheses can be generated for different fetal fractions. For each case, an expected distribution of the proportion of corrected reads can be calculated given expected statistical variation in the selection and measurement of the various DNA loci. The observed corrected proportion of reads can be compared to the distribution of the expected proportion of corrected reads, and a likelihood ratio can be calculated for the disomy and trisomy hypotheses, for each of the samples of unknown ploidy. The ploidy state associated with the hypothesis with the highest calculated likelihood can be selected as the correct ploidy state.

**[284]** In some embodiments, a subset of the samples with a sufficiently low likelihood of having cancer may be selected to act as a control set of samples. The subset can be a fixed number, or it can be a variable number that is based on choosing only those samples that fall below a threshold. The quantitative data from the subset of samples may be combined, averaged, or combined using a weighted average where the weighting is based on the likelihood of the sample being normal. The quantitative data may be used to determine the per-locus bias for the amplification the sequencing of samples in the instant batch of

control samples. The per-locus bias may also include data from other batches of samples. The per-locus bias may indicate the relative over- or under-amplification that is observed for that locus compared to other loci, making the assumption that the subset of samples do not contain any CNVs, and that any observed over or under-amplification is due to amplification and/or sequencing or other bias. The per-locus bias may take into account the GC content of the amplicon. The loci may be grouped into groups of loci for the purpose of calculating a per-locus bias. Once the per-locus bias has been calculated for each locus in the plurality of loci, the sequencing data for one or more of the samples that are not in the subset of the samples, and optionally one or more of the samples that are in the subset of samples, may be corrected by adjusting the quantitative measurements for each locus to remove the effect of the bias at that locus. For example, if SNP 1 was observed, in the subset of patients, to have a depth of read that is twice as great as the average, the adjustment may involve replacing the number of reads corresponding from SNP 1 with a number that is half as great. If the locus in question is a SNP, the adjustment may involve cutting the number of reads corresponding to each of the alleles at that locus in half. Once the sequencing data for each of the loci in one or more samples has been adjusted, it may be analyzed using a method for the purpose of detecting the presence of a CNV at one or more chromosomal regions.

**[285]** In an example, sample A is a mixture of amplified DNA originating from a mixture of normal and cancerous cells that is analyzed using a quantitative method. The following illustrates exemplary possible data. A region of the q arm on chromosome 22 is found to only have 90% as much DNA mapping to that region as expected; a focal region corresponding to the HER2 gene is found to have 150% as much DNA mapping to that region as expected; and the p-arm of chromosome 5 is found to have 105% as much DNA mapping to it as expected. A clinician may infer that the sample has a deletion of a region on the q arm on chromosome 22, and a duplication of the HER2 gene. The clinician may infer that since the 22q deletions are common in breast cancer, and that since cells with a deletion of the 22q region on both chromosomes usually do not survive, that approximately 20% of the DNA in the sample came from cells with a 22q deletion on one of the two chromosomes. The clinician may also infer that if the DNA from the mixed sample that originated from tumor cells originated from a set of genetically tumor cells whose HER2 region and 22q regions were homogenous, then the cells contained a five-fold duplication of the HER2 region.

**[286]** In an example, Sample A is also analyzed using an allelic method. The following illustrates exemplary possible data. The two haplotypes on same region on the q arm on chromosome 22 are present in a ratio of 4:5; the two haplotypes in a focal region corresponding to the HER2 gene are

control samples. The per-locus bias may also include data from other batches of samples. The per-locus bias may indicate the relative over- or under-amplification that is observed for that locus compared to other loci, making the assumption that the subset of samples do not contain any CNVs, and that any observed over or under-amplification is due to amplification and/or sequencing or other bias. The per-locus bias may take into account the GC content of the amplicon. The loci may be grouped into groups of loci for the purpose of calculating a per-locus bias. Once the per-locus bias has been calculated for each locus in the plurality of loci, the sequencing data for one or more of the samples that are not in the subset of the samples, and optionally one or more of the samples that are in the subset of samples, may be corrected by adjusting the quantitative measurements for each locus to remove the effect of the bias at that locus. For example, if SNP 1 was observed, in the subset of patients, to have a depth of read that is twice as great as the average, the adjustment may involve replacing the number of reads corresponding from SNP 1 with a number that is half as great. If the locus in question is a SNP, the adjustment may involve cutting the number of reads corresponding to each of the alleles at that locus in half. Once the sequencing data for each of the loci in one or more samples has been adjusted, it may be analyzed using a method for the purpose of detecting the presence of a CNV at one or more chromosomal regions.

**[285]** In an example, sample A is a mixture of amplified DNA originating from a mixture of normal and cancerous cells that is analyzed using a quantitative method. The following illustrates exemplary possible data. A region of the q arm on chromosome 22 is found to only have 90% as much DNA mapping to that region as expected; a focal region corresponding to the HER2 gene is found to have 150% as much DNA mapping to that region as expected; and the p-arm of chromosome 5 is found to have 105% as much DNA mapping to it as expected. A clinician may infer that the sample has a deletion of a region on the q arm on chromosome 22, and a duplication of the HER2 gene. The clinician may infer that since the 22q deletions are common in breast cancer, and that since cells with a deletion of the 22q region on both chromosomes usually do not survive, that approximately 20% of the DNA in the sample came from cells with a 22q deletion on one of the two chromosomes. The clinician may also infer that if the DNA from the mixed sample that originated from tumor cells originated from a set of genetically tumor cells whose HER2 region and 22q regions were homogenous, then the cells contained a five-fold duplication of the HER2 region.

**[286]** In an example, Sample A is also analyzed using an allelic method. The following illustrates exemplary possible data. The two haplotypes on same region on the q arm on chromosome 22 are present in a ratio of 4:5; the two haplotypes in a focal region corresponding to the HER2 gene are

present in ratios of 1:2; and the two haplotypes in the p-arm of chromosome 5 are present in ratios of 20:21. All other assayed regions of the genome have no statistically significant excess of either haplotype. A clinician may infer that the sample contains DNA from a tumor with a CNV in the 22q region, the HER2 region, and the 5p arm. Based on the knowledge that 22q deletions are very common in breast cancer, and/or the quantitative analysis showing an under-representation of the amount of DNA mapping to the 22q region of the genome, the clinician may infer the existence of a tumor with a 22q deletion. Based on the knowledge that HER2 amplifications are very common in breast cancer, and/or the quantitative analysis showing an over-representation of the amount of DNA mapping to the HER2 region of the genome, the clinician may infer the existence of a tumor with a HER2 amplification.

#### Exemplary Reference Chromosomes or Chromosome Segments

**[287]** In some embodiments, any of the methods described herein are also performed on one or more reference chromosomes or chromosomes segments and the results are compared to those for one or more chromosomes or chromosome segments of interest.

**[288]** In some embodiments, the reference chromosome or chromosome segment is used as a control for what would be expected for the absence of a CNV. In some embodiments, the reference is the same chromosome or chromosome segment from one or more different samples known or expected to not have a deletion or duplication in that chromosome or chromosome segment. In some embodiments, the reference is a different chromosome or chromosome segment from the sample being tested that is expected to be disomic. In some embodiments, the reference is a different segment from one of the chromosomes of interest in the same sample that is being tested. For example, the reference may be one or more segments outside of the region of a potential deletion or duplication. Having a reference on the same chromosome that is being tested avoids variability between different chromosomes, such as differences in metabolism, apoptosis, histones, inactivation, and/or amplification between chromosomes. Analyzing segments without a CNV on the same chromosome as the one being tested can also be used to determine differences in metabolism, apoptosis, histones, inactivation, and/or amplification between homologs, allowing the level of variability between homologs in the absence of a CNV to be determined for comparison to the results from a potential CNV. In some embodiments, the magnitude of the difference between the calculated and expected allele ratios for a potential CNV is greater than the corresponding magnitude for the reference, thereby confirming the presence of a CNV.

**[289]** In some embodiments, the reference chromosome or chromosome segment is used as a control for what would be expected for the presence of a CNV, such as a particular deletion or duplication of

present in ratios of 1:2; and the two haplotypes in the p-arm of chromosome 5 are present in ratios of 20:21. All other assayed regions of the genome have no statistically significant excess of either haplotype. A clinician may infer that the sample contains DNA from a tumor with a CNV in the 22q region, the HER2 region, and the 5p arm. Based on the knowledge that 22q deletions are very common in breast cancer, and/or the quantitative analysis showing an under-representation of the amount of DNA mapping to the 22q region of the genome, the clinician may infer the existence of a tumor with a 22q deletion. Based on the knowledge that HER2 amplifications are very common in breast cancer, and/or the quantitative analysis showing an over-representation of the amount of DNA mapping to the HER2 region of the genome, the clinician may infer the existence of a tumor with a HER2 amplification.

#### Exemplary Reference Chromosomes or Chromosome Segments

**[287]** In some embodiments, any of the methods described herein are also performed on one or more reference chromosomes or chromosomes segments and the results are compared to those for one or more chromosomes or chromosome segments of interest.

**[288]** In some embodiments, the reference chromosome or chromosome segment is used as a control for what would be expected for the absence of a CNV. In some embodiments, the reference is the same chromosome or chromosome segment from one or more different samples known or expected to not have a deletion or duplication in that chromosome or chromosome segment. In some embodiments, the reference is a different chromosome or chromosome segment from the sample being tested that is expected to be disomic. In some embodiments, the reference is a different segment from one of the chromosomes of interest in the same sample that is being tested. For example, the reference may be one or more segments outside of the region of a potential deletion or duplication. Having a reference on the same chromosome that is being tested avoids variability between different chromosomes, such as differences in metabolism, apoptosis, histones, inactivation, and/or amplification between chromosomes. Analyzing segments without a CNV on the same chromosome as the one being tested can also be used to determine differences in metabolism, apoptosis, histones, inactivation, and/or amplification between homologs, allowing the level of variability between homologs in the absence of a CNV to be determined for comparison to the results from a potential CNV. In some embodiments, the magnitude of the difference between the calculated and expected allele ratios for a potential CNV is greater than the corresponding magnitude for the reference, thereby confirming the presence of a CNV.

**[289]** In some embodiments, the reference chromosome or chromosome segment is used as a control for what would be expected for the presence of a CNV, such as a particular deletion or duplication of

interest. In some embodiments, the reference is the same chromosome or chromosome segment from one or more different samples known or expected to have a deletion or duplication in that chromosome or chromosome segment. In some embodiments, the reference is a different chromosome or chromosome segment from the sample being tested that is known or expected to have a CNV. In some embodiments, the magnitude of the difference between the calculated and expected allele ratios for a potential CNV is similar to (such as not significantly different) than the corresponding magnitude for the reference for the CNV, thereby confirming the presence of a CNV. In some embodiments, the magnitude of the difference between the calculated and expected allele ratios for a potential CNV is less than (such as significantly less) than the corresponding magnitude for the reference for the CNV, thereby confirming the absence of a CNV. In some embodiments, one or more loci for which the genotype of a cancer cell (or DNA or RNA from a cancer cell such as cfDNA or cfRNA) differs from the genotype of a noncancerous cell (or DNA or RNA from a noncancerous cell such as cfDNA or cfRNA) is used to determine the tumor fraction. The tumor fraction can be used to determine whether the overrepresentation of the number of copies of the first homologous chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment. The tumor fraction can also be used to determine the number of extra copies of a chromosome segment or chromosome that is duplicated (such as whether there are 1, 2, 3, 4, or more extra copies), such as to differentiate a sample with four extra chromosome copies and a tumor fraction of 10% from a sample with two extra chromosome copies and a tumor fraction of 20%. The tumor fraction can also be used to determine how well the observed data fits the expected data for possible CNVs. In some embodiments, the degree of overrepresentation of a CNV is used to select a particular therapy or therapeutic regimen for the individual. For example, some therapeutic agents are only effective for at least four, six, or more copies of a chromosome segment.

**[290]** In some embodiments, the one or more loci used to determine the tumor fraction are on a reference chromosome or chromosomes segment, such as a chromosome or chromosome segment known or expected to be disomic, a chromosome or chromosome segment that is rarely duplicated or deleted in cancer cells in general or in a particular type of cancer that an individual is known to have or is at increased risk of having, or a chromosome or chromosome segment that is unlikely to be aneuploid (such segment that is expected to lead to cell death if deleted or duplicated). In some embodiments, any of the methods of the invention are used to confirm that the reference chromosome or chromosome segment is disomic in both the cancer cells and noncancerous cells. In some embodiments, one or more chromosomes or chromosomes segments for which the confidence for a disomy call is high are used.

interest. In some embodiments, the reference is the same chromosome or chromosome segment from one or more different samples known or expected to have a deletion or duplication in that chromosome or chromosome segment. In some embodiments, the reference is a different chromosome or chromosome segment from the sample being tested that is known or expected to have a CNV. In some embodiments, the magnitude of the difference between the calculated and expected allele ratios for a potential CNV is similar to (such as not significantly different) than the corresponding magnitude for the reference for the CNV, thereby confirming the presence of a CNV. In some embodiments, the magnitude of the difference between the calculated and expected allele ratios for a potential CNV is less than (such as significantly less) than the corresponding magnitude for the reference for the CNV, thereby confirming the absence of a CNV. In some embodiments, one or more loci for which the genotype of a cancer cell (or DNA or RNA from a cancer cell such as cfDNA or cfRNA) differs from the genotype of a noncancerous cell (or DNA or RNA from a noncancerous cell such as cfDNA or cfRNA) is used to determine the tumor fraction. The tumor fraction can be used to determine whether the overrepresentation of the number of copies of the first homologous chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment. The tumor fraction can also be used to determine the number of extra copies of a chromosome segment or chromosome that is duplicated (such as whether there are 1, 2, 3, 4, or more extra copies), such as to differentiate a sample with four extra chromosome copies and a tumor fraction of 10% from a sample with two extra chromosome copies and a tumor fraction of 20%. The tumor fraction can also be used to determine how well the observed data fits the expected data for possible CNVs. In some embodiments, the degree of overrepresentation of a CNV is used to select a particular therapy or therapeutic regimen for the individual. For example, some therapeutic agents are only effective for at least four, six, or more copies of a chromosome segment.

**[290]** In some embodiments, the one or more loci used to determine the tumor fraction are on a reference chromosome or chromosomes segment, such as a chromosome or chromosome segment known or expected to be disomic, a chromosome or chromosome segment that is rarely duplicated or deleted in cancer cells in general or in a particular type of cancer that an individual is known to have or is at increased risk of having, or a chromosome or chromosome segment that is unlikely to be aneuploid (such segment that is expected to lead to cell death if deleted or duplicated). In some embodiments, any of the methods of the invention are used to confirm that the reference chromosome or chromosome segment is disomic in both the cancer cells and noncancerous cells. In some embodiments, one or more chromosomes or chromosomes segments for which the confidence for a disomy call is high are used.

**[291]** Exemplary loci that can be used to determine the tumor fraction include polymorphisms or mutations (such as SNPs) in a cancer cell (or DNA or RNA such as cfDNA or cfRNA from a cancer cell) that aren't present in a noncancerous cell (or DNA or RNA from a noncancerous cell) in the individual. In some embodiments, the tumor fraction is determined by identifying those polymorphic loci where a cancer cell (or DNA or RNA from a cancer cell) has an allele that is absent in noncancerous cells (or DNA or RNA from a noncancerous cell) in a sample (such as a plasma sample or tumor biopsy) from an individual; and using the amount of the allele unique to the cancer cell at one or more of the identified polymorphic loci to determine the tumor fraction in the sample. In some embodiments, a noncancerous cell is homozygous for a first allele at the polymorphic locus, and a cancer cell is (i) heterozygous for the first allele and a second allele or (ii) homozygous for a second allele at the polymorphic locus. In some embodiments, a noncancerous cell is heterozygous for a first allele and a second allele at the polymorphic locus, and a cancer cell is (i) has one or two copies of a third allele at the polymorphic locus. In some embodiments, the cancer cells are assumed or known to only have one copy of the allele that is not present in the noncancerous cells. For example, if the genotype of the noncancerous cells is AA and the cancer cells is AB and 5% of the signal at that locus in a sample is from the B allele and 95% is from the A allele, then the tumor fraction of the sample is 10%. In some embodiments, the cancer cells are assumed or known to have two copies of the allele that is not present in the noncancerous cells. For example, if the genotype of the noncancerous cells is AA and the cancer cells is BB and 5% of the signal at that locus in a sample is from the B allele and 95% is from the A allele, the tumor fraction of the sample is 5%. In some embodiments, multiple loci for which the cancer cells have an allele not in the noncancerous cells are analyzed to determine which of the loci in the cancer cells are heterozygous and which are homozygous. For example for loci in which the noncancerous cells are AA, if the signal from the B allele is -5% at some loci and -10% at some loci, then the cancer cells are assumed to be heterozygous at loci with ~5% B allele, and homozygous at loci with -10% B allele (indicating the tumor fraction is -10%).

**[292]** Exemplary loci that can be used to determine the tumor fraction include loci for which a cancer cell and noncancerous cell have one allele in common (such as loci in which the cancer cell is AB and the noncancerous cell is BB, or the cancer cell is BB and the noncancerous cell is AB). The amount of A signal, the amount of B signal, or the ratio of A to B signal in a mixed sample (containing DNA or RNA from a cancer cell and a noncancerous cell) is compared to the corresponding value for (i) a sample containing DNA or RNA from only cancer cells or (ii) a sample containing DNA or RNA from only noncancerous cells. The difference in values is used to determine the tumor fraction of the mixed sample.

**[291]** Exemplary loci that can be used to determine the tumor fraction include polymorphisms or mutations (such as SNPs) in a cancer cell (or DNA or RNA such as cfDNA or cfRNA from a cancer cell) that aren't present in a noncancerous cell (or DNA or RNA from a noncancerous cell) in the individual. In some embodiments, the tumor fraction is determined by identifying those polymorphic loci where a cancer cell (or DNA or RNA from a cancer cell) has an allele that is absent in noncancerous cells (or DNA or RNA from a noncancerous cell) in a sample (such as a plasma sample or tumor biopsy) from an individual; and using the amount of the allele unique to the cancer cell at one or more of the identified polymorphic loci to determine the tumor fraction in the sample. In some embodiments, a noncancerous cell is homozygous for a first allele at the polymorphic locus, and a cancer cell is (i) heterozygous for the first allele and a second allele or (ii) homozygous for a second allele at the polymorphic locus. In some embodiments, a noncancerous cell is heterozygous for a first allele and a second allele at the polymorphic locus, and a cancer cell is (i) has one or two copies of a third allele at the polymorphic locus. In some embodiments, the cancer cells are assumed or known to only have one copy of the allele that is not present in the noncancerous cells. For example, if the genotype of the noncancerous cells is AA and the cancer cells is AB and 5% of the signal at that locus in a sample is from the B allele and 95% is from the A allele, then the tumor fraction of the sample is 10%. In some embodiments, the cancer cells are assumed or known to have two copies of the allele that is not present in the noncancerous cells. For example, if the genotype of the noncancerous cells is AA and the cancer cells is BB and 5% of the signal at that locus in a sample is from the B allele and 95% is from the A allele, the tumor fraction of the sample is 5%. In some embodiments, multiple loci for which the cancer cells have an allele not in the noncancerous cells are analyzed to determine which of the loci in the cancer cells are heterozygous and which are homozygous. For example for loci in which the noncancerous cells are AA, if the signal from the B allele is -5% at some loci and -10% at some loci, then the cancer cells are assumed to be heterozygous at loci with ~5% B allele, and homozygous at loci with -10% B allele (indicating the tumor fraction is -10%).

**[292]** Exemplary loci that can be used to determine the tumor fraction include loci for which a cancer cell and noncancerous cell have one allele in common (such as loci in which the cancer cell is AB and the noncancerous cell is BB, or the cancer cell is BB and the noncancerous cell is AB). The amount of A signal, the amount of B signal, or the ratio of A to B signal in a mixed sample (containing DNA or RNA from a cancer cell and a noncancerous cell) is compared to the corresponding value for (i) a sample containing DNA or RNA from only cancer cells or (ii) a sample containing DNA or RNA from only noncancerous cells. The difference in values is used to determine the tumor fraction of the mixed sample.

**[293]** In some embodiments, loci that can be used to determine the tumor fraction are selected based on the genotype of (i) a sample containing DNA or RNA from only cancer cells, and/or (ii) a sample containing DNA or RNA from only noncancerous cells. In some embodiments, the loci are selected based on analysis of the mixed sample, such as loci for which the absolute or relative amounts of each allele differs from what would be expected if both the cancer and noncancerous cells have the same genotype at a particular locus. For example, if the cancer and noncancerous cells have the same genotype, the loci would be expected to produce 0% B signal if all the cells are AA, 50% B signal if all the cells are AB, or 100% B signal if all the cells are BB. Other values for the B signal indicate that the genotype of the cancer and noncancerous cells are different at that locus and thus that locus can be used to determine the tumor fraction.

**[294]** In some embodiments, the tumor fraction calculated based on the alleles at one or more loci is compared to the tumor fraction calculated using one or more of the counting methods disclosed herein.

#### Exemplary Methods for Detecting a Phenotype or Analyzing Multiple Mutations

**[295]** In some embodiments, the method includes analyzing a sample for a set of mutations associated with a disease or disorder (such as cancer) or an increased risk for a disease or disorder. There are strong correlations between events within classes (such as M or C cancer classes) which can be used to improve the signal to noise ratio of a method and classify tumors into distinct clinical subsets. For example, borderline results for a few mutations (such as a few CNVs) on one or more chromosomes or chromosomes segments considered jointly may be a very strong signal. In some embodiments, determining the presence or absence of multiple polymorphisms or mutations of interest (such as 2, 3, 4, 5, 8, 10, 12, 15, or more) increases the sensitivity and/or specificity of the determination of the presence or absence of a disease or disorder such as cancer, or an increased risk for with a disease or disorder such as cancer. In some embodiments, the correlation between events across multiple chromosomes is used to more powerfully look at a signal compared to looking at each of them individually. The design of the method itself can be optimized to best categorize tumors. This may be incredibly useful for early detection and screening—vis-a-vis recurrence where sensitivity to one particular mutation/CNV may be paramount. In some embodiments, the events are not always correlated but have a probability of being correlated. In some embodiments, a matrix estimation formulation with a noise covariance matrix that has off diagonal terms is used.

**[296]** In some embodiments, the invention features a method for detecting a phenotype (such as a cancer phenotype) in an individual, wherein the phenotype is defined by the presence of at least one of a

**[293]** In some embodiments, loci that can be used to determine the tumor fraction are selected based on the genotype of (i) a sample containing DNA or RNA from only cancer cells, and/or (ii) a sample containing DNA or RNA from only noncancerous cells. In some embodiments, the loci are selected based on analysis of the mixed sample, such as loci for which the absolute or relative amounts of each allele differs from what would be expected if both the cancer and noncancerous cells have the same genotype at a particular locus. For example, if the cancer and noncancerous cells have the same genotype, the loci would be expected to produce 0% B signal if all the cells are AA, 50% B signal if all the cells are AB, or 100% B signal if all the cells are BB. Other values for the B signal indicate that the genotype of the cancer and noncancerous cells are different at that locus and thus that locus can be used to determine the tumor fraction.

**[294]** In some embodiments, the tumor fraction calculated based on the alleles at one or more loci is compared to the tumor fraction calculated using one or more of the counting methods disclosed herein.

#### Exemplary Methods for Detecting a Phenotype or Analyzing Multiple Mutations

**[295]** In some embodiments, the method includes analyzing a sample for a set of mutations associated with a disease or disorder (such as cancer) or an increased risk for a disease or disorder. There are strong correlations between events within classes (such as M or C cancer classes) which can be used to improve the signal to noise ratio of a method and classify tumors into distinct clinical subsets. For example, borderline results for a few mutations (such as a few CNVs) on one or more chromosomes or chromosomes segments considered jointly may be a very strong signal. In some embodiments, determining the presence or absence of multiple polymorphisms or mutations of interest (such as 2, 3, 4, 5, 8, 10, 12, 15, or more) increases the sensitivity and/or specificity of the determination of the presence or absence of a disease or disorder such as cancer, or an increased risk for with a disease or disorder such as cancer. In some embodiments, the correlation between events across multiple chromosomes is used to more powerfully look at a signal compared to looking at each of them individually. The design of the method itself can be optimized to best categorize tumors. This may be incredibly useful for early detection and screening—vis-a-vis recurrence where sensitivity to one particular mutation/CNV may be paramount. In some embodiments, the events are not always correlated but have a probability of being correlated. In some embodiments, a matrix estimation formulation with a noise covariance matrix that has off diagonal terms is used.

**[296]** In some embodiments, the invention features a method for detecting a phenotype (such as a cancer phenotype) in an individual, wherein the phenotype is defined by the presence of at least one of a

set of mutations. In some embodiments, the method includes obtaining DNA or RNA measurements for a sample of DNA or RNA from one or more cells from the individual, wherein one or more of the cells is suspected of having the phenotype; and analyzing the DNA or RNA measurements to determine, for each of the mutations in the set of mutations, the likelihood that at least one of the cells has that mutation. In some embodiments, the method includes determining that the individual has the phenotype if either (i) for at least one of the mutations, the likelihood that at least one of the cells contains that mutations is greater than a threshold, or (ii) for at least one of the mutations, the likelihood that at least one of the cells has that mutations is less than the threshold, and for a plurality of the mutations, the combined likelihood that at least one of the cells has at least one of the mutations is greater than the threshold. In some embodiments, one or more cells have a subset or all of the mutations in the set of mutations. In some embodiments, the subset of mutations is associated with cancer or an increased risk for cancer. In some embodiments, the set of mutations includes a subset or all of the mutations in the M class of cancer mutations (Ciriello, Nat Genet. 45(10): 1127-1133, 2013, doi: 10.1038/ng.2762, which is hereby incorporated by reference in its entirety). In some embodiments, the set of mutations includes a subset or all of the mutations in the C class of cancer mutations (Ciriello, *supra*). In some embodiments, the sample includes cell-free DNA or RNA. In some embodiments, the DNA or RNA measurements include measurements (such as the quantity of each allele at each locus) at a set of polymorphic loci on one or more chromosomes or chromosome segments of interest.

#### Exemplary Methods for Paternity Testing or Genetic Relatedness Testing

[297] The methods of the invention can be used to improve the accuracy of paternity testing or other genetic relatedness testing (see, *e.g.*, U.S. Publication No. 2012/0122701, filed December 22, 2011, which is hereby incorporated by reference in its entirety). For example, the multiplex PCR method can allow thousands of polymorphic loci (such as SNPs) to be analyzed for use in the PARENTAL SUPPORT algorithm described herein to determine whether an alleged father is the biological father of a fetus. In some embodiments, the invention features a method for establishing whether an alleged father is the biological father of a fetus that is gestating in a pregnant mother. In some embodiments, the method involves obtaining phased genetic data for the alleged father (such as by using another of the methods described herein for phasing genetic data), wherein the phased genetic data comprises the identity of the allele present for each locus in a set of polymorphic loci on a first homologous chromosome segment and a second homologous chromosome segment in the alleged father. In some embodiments, the method involves obtaining genetic data at the set of polymorphic loci on the chromosome or chromosome

set of mutations. In some embodiments, the method includes obtaining DNA or RNA measurements for a sample of DNA or RNA from one or more cells from the individual, wherein one or more of the cells is suspected of having the phenotype; and analyzing the DNA or RNA measurements to determine, for each of the mutations in the set of mutations, the likelihood that at least one of the cells has that mutation. In some embodiments, the method includes determining that the individual has the phenotype if either (i) for at least one of the mutations, the likelihood that at least one of the cells contains that mutations is greater than a threshold, or (ii) for at least one of the mutations, the likelihood that at least one of the cells has that mutations is less than the threshold, and for a plurality of the mutations, the combined likelihood that at least one of the cells has at least one of the mutations is greater than the threshold. In some embodiments, one or more cells have a subset or all of the mutations in the set of mutations. In some embodiments, the subset of mutations is associated with cancer or an increased risk for cancer. In some embodiments, the set of mutations includes a subset or all of the mutations in the M class of cancer mutations (Ciriello, Nat Genet. 45(10): 1127-1133, 2013, doi: 10.1038/ng.2762, which is hereby incorporated by reference in its entirety). In some embodiments, the set of mutations includes a subset or all of the mutations in the C class of cancer mutations (Ciriello, *supra*). In some embodiments, the sample includes cell-free DNA or RNA. In some embodiments, the DNA or RNA measurements include measurements (such as the quantity of each allele at each locus) at a set of polymorphic loci on one or more chromosomes or chromosome segments of interest.

#### Exemplary Methods for Paternity Testing or Genetic Relatedness Testing

[297] The methods of the invention can be used to improve the accuracy of paternity testing or other genetic relatedness testing (see, *e.g.*, U.S. Publication No. 2012/0122701, filed December 22, 2011, which is hereby incorporated by reference in its entirety). For example, the multiplex PCR method can allow thousands of polymorphic loci (such as SNPs) to be analyzed for use in the PARENTAL SUPPORT algorithm described herein to determine whether an alleged father is the biological father of a fetus. In some embodiments, the invention features a method for establishing whether an alleged father is the biological father of a fetus that is gestating in a pregnant mother. In some embodiments, the method involves obtaining phased genetic data for the alleged father (such as by using another of the methods described herein for phasing genetic data), wherein the phased genetic data comprises the identity of the allele present for each locus in a set of polymorphic loci on a first homologous chromosome segment and a second homologous chromosome segment in the alleged father. In some embodiments, the method involves obtaining genetic data at the set of polymorphic loci on the chromosome or chromosome

segment in a mixed sample of DNA comprising fetal DNA and maternal DNA from the mother of the fetus by measuring the quantity of each allele at each locus. In some embodiments, the method involves calculating, on a computer, expected genetic data for the mixed sample of DNA from the phased genetic data for the alleged father; determining, on a computer, the probability that the alleged father is the biological father of the fetus by comparing the obtaining genetic data made on the mixed sample of DNA to the expected genetic data for the mixed sample of DNA; and establishing whether the alleged father is the biological father of the fetus using the determined probability that the alleged father is the biological father of the fetus. In some embodiments, the method involves obtaining phased genetic data for the biological mother of the fetus (such as by using another of the methods described herein for phasing genetic data), wherein the phased genetic data comprises the identity of the allele present for each locus in a set of polymorphic loci on a first homologous chromosome segment and a second homologous chromosome segment in the mother. In some embodiments, the method involves obtaining phased genetic data for the fetus (such as by using another of the methods described herein for phasing genetic data), wherein the phased genetic data comprises the identity of the allele present for each locus in a set of polymorphic loci on a first homologous chromosome segment and a second homologous chromosome segment in the fetus. In some embodiments, the method involves calculating, on a computer, expected genetic data for the mixed sample of DNA using the phased genetic data for the alleged father and using the phased genetic data for the mother and/or the phased genetic data for the fetus.

**[298]** In some embodiments, the invention features a method for establishing whether an alleged father is the biological father of a fetus that is gestating in a pregnant mother. In some embodiments, the method involves obtaining phased genetic data for the alleged father (such as by using another of the methods described herein for phasing genetic data), wherein the phased genetic data comprises the identity of the allele present for each locus in a set of polymorphic loci on a first homologous chromosome segment and a second homologous chromosome segment in the alleged father. In some embodiments, the method involves obtaining genetic data at the set of polymorphic loci on the chromosome or chromosome segment in a mixed sample of DNA comprising fetal DNA and maternal DNA from the mother of the fetus by measuring the quantity of each allele at each locus. In some embodiments, the method involves identifying (i) alleles that are present in the fetal DNA but are absent in the maternal DNA at polymorphic loci, and/or identifying (i) alleles that are absent in the fetal DNA and the maternal DNA at polymorphic loci. In some embodiments, the method involves determining, on a computer, the probability that the alleged father is the biological father of the fetus; wherein the determination comprises: (1) comparing (i)

segment in a mixed sample of DNA comprising fetal DNA and maternal DNA from the mother of the fetus by measuring the quantity of each allele at each locus. In some embodiments, the method involves calculating, on a computer, expected genetic data for the mixed sample of DNA from the phased genetic data for the alleged father; determining, on a computer, the probability that the alleged father is the biological father of the fetus by comparing the obtaining genetic data made on the mixed sample of DNA to the expected genetic data for the mixed sample of DNA; and establishing whether the alleged father is the biological father of the fetus using the determined probability that the alleged father is the biological father of the fetus. In some embodiments, the method involves obtaining phased genetic data for the biological mother of the fetus (such as by using another of the methods described herein for phasing genetic data), wherein the phased genetic data comprises the identity of the allele present for each locus in a set of polymorphic loci on a first homologous chromosome segment and a second homologous chromosome segment in the mother. In some embodiments, the method involves obtaining phased genetic data for the fetus (such as by using another of the methods described herein for phasing genetic data), wherein the phased genetic data comprises the identity of the allele present for each locus in a set of polymorphic loci on a first homologous chromosome segment and a second homologous chromosome segment in the fetus. In some embodiments, the method involves calculating, on a computer, expected genetic data for the mixed sample of DNA using the phased genetic data for the alleged father and using the phased genetic data for the mother and/or the phased genetic data for the fetus.

**[298]** In some embodiments, the invention features a method for establishing whether an alleged father is the biological father of a fetus that is gestating in a pregnant mother. In some embodiments, the method involves obtaining phased genetic data for the alleged father (such as by using another of the methods described herein for phasing genetic data), wherein the phased genetic data comprises the identity of the allele present for each locus in a set of polymorphic loci on a first homologous chromosome segment and a second homologous chromosome segment in the alleged father. In some embodiments, the method involves obtaining genetic data at the set of polymorphic loci on the chromosome or chromosome segment in a mixed sample of DNA comprising fetal DNA and maternal DNA from the mother of the fetus by measuring the quantity of each allele at each locus. In some embodiments, the method involves identifying (i) alleles that are present in the fetal DNA but are absent in the maternal DNA at polymorphic loci, and/or identifying (i) alleles that are absent in the fetal DNA and the maternal DNA at polymorphic loci. In some embodiments, the method involves determining, on a computer, the probability that the alleged father is the biological father of the fetus; wherein the determination comprises: (1) comparing (i)

the alleles that are present in the fetal DNA but are absent in the maternal DNA at polymorphic loci to (ii) the alleles at the corresponding polymorphic loci in the genetic material from the alleged father, and/or (2) comparing (i) the alleles that are absent in the fetal DNA and the maternal DNA at polymorphic loci to (ii) the alleles at the corresponding polymorphic loci in the genetic material from the alleged father; and establishing whether the alleged father is the biological father of the fetus using the determined probability that the alleged father is the biological father of the fetus.

**[299]** In some embodiments, a method described above for determining whether an alleged father is the biological father of the fetus is used to determine if an alleged relative (such as a grandparent, sibling, aunt, or uncle) of a fetus is an actual biological relative of the fetus (such as by using genetic data of the alleged relative instead of genetic data of the alleged father).

#### Exemplary Combinations of Methods

**[300]** To increase the accuracy of the results, two or more methods (such as any of the methods of the invention or any known method) for detecting the presence or absence of a CNV are performed. In some embodiments, one or more methods for analyzing a factor (such as any of the method described herein or any known method) indicative of the presence or absence of a disease or disorder or an increased risk for a disease or disorder are performed.

**[301]** In some embodiments, standard mathematical techniques are used to calculate the covariance and/or correlation between two or more methods. Standard mathematical techniques may also be used to determine the combined probability of a particular hypothesis based on two or more tests. Exemplary techniques include meta-analysis, Fisher's combined probability test for independent tests, Brown's method for combining dependent p-values with known covariance, and Kost's method for combining dependent p-values with unknown covariance. In cases where the likelihoods are determined by a first method in a way that is orthogonal, or unrelated, to the way in which a likelihood is determined for a second method, combining the likelihoods is straightforward and can be done by multiplication and normalization, or by using a formula such as:

$$R_{comb} = R_1 R_2 / [R_1 R_2 + (1 - R_1)(1 - R_2)]$$

$R_{comb}$  is the combined likelihood, and  $R_1$  and  $R_2$  are the individual likelihoods. For example, if the likelihood of trisomy from method 1 is 90%, and the likelihood of trisomy from method 2 is 95%, then combining the outputs from the two methods allows the clinician to conclude that the fetus is trisomic with a likelihood of  $(0.90)(0.95) / [(0.90)(0.95) + (1 - 0.90)(1 - 0.95)] = 99.42\%$ . In cases where the first and the second methods are not orthogonal, that is, where there is a correlation between the two methods,

the alleles that are present in the fetal DNA but are absent in the maternal DNA at polymorphic loci to (ii) the alleles at the corresponding polymorphic loci in the genetic material from the alleged father, and/or (2) comparing (i) the alleles that are absent in the fetal DNA and the maternal DNA at polymorphic loci to (ii) the alleles at the corresponding polymorphic loci in the genetic material from the alleged father; and establishing whether the alleged father is the biological father of the fetus using the determined probability that the alleged father is the biological father of the fetus.

**[299]** In some embodiments, a method described above for determining whether an alleged father is the biological father of the fetus is used to determine if an alleged relative (such as a grandparent, sibling, aunt, or uncle) of a fetus is an actual biological relative of the fetus (such as by using genetic data of the alleged relative instead of genetic data of the alleged father).

#### Exemplary Combinations of Methods

**[300]** To increase the accuracy of the results, two or more methods (such as any of the methods of the invention or any known method) for detecting the presence or absence of a CNV are performed. In some embodiments, one or more methods for analyzing a factor (such as any of the method described herein or any known method) indicative of the presence or absence of a disease or disorder or an increased risk for a disease or disorder are performed.

**[301]** In some embodiments, standard mathematical techniques are used to calculate the covariance and/or correlation between two or more methods. Standard mathematical techniques may also be used to determine the combined probability of a particular hypothesis based on two or more tests. Exemplary techniques include meta-analysis, Fisher's combined probability test for independent tests, Brown's method for combining dependent p-values with known covariance, and Kost's method for combining dependent p-values with unknown covariance. In cases where the likelihoods are determined by a first method in a way that is orthogonal, or unrelated, to the way in which a likelihood is determined for a second method, combining the likelihoods is straightforward and can be done by multiplication and normalization, or by using a formula such as:

$$R_{comb} = R_1 R_2 / [R_1 R_2 + (1 - R_1)(1 - R_2)]$$

$R_{comb}$  is the combined likelihood, and  $R_1$  and  $R_2$  are the individual likelihoods. For example, if the likelihood of trisomy from method 1 is 90%, and the likelihood of trisomy from method 2 is 95%, then combining the outputs from the two methods allows the clinician to conclude that the fetus is trisomic with a likelihood of  $(0.90)(0.95) / [(0.90)(0.95) + (1 - 0.90)(1 - 0.95)] = 99.42\%$ . In cases where the first and the second methods are not orthogonal, that is, where there is a correlation between the two methods,

the likelihoods can still be combined.

**[302]** Exemplary methods of analyzing multiple factors or variables are disclosed in U.S. Patent No. 8,024,128 issued on September 20, 2011; U.S. Publication No. 2007/0027636, filed July 31, 2006; and U.S. Publication No. 2007/0178501, filed December 6, 2006, which are each hereby incorporated by reference in its entirety).

**[303]** In various embodiments, the combined probability of a particular hypothesis or diagnosis is greater than 80, 85, 90, 92, 94, 96, 98, 99, or 99.9%, or is greater than some other threshold value.

#### Limit of Detection

**[304]** In some embodiments, a limit of detection of a mutation (such as an SNV or CNV) of a method of the invention is less than or equal to 10, 5, 2, 1, 0.5, 0.1, 0.05, 0.01, or 0.005%. In some embodiments, a limit of detection of a mutation (such as an SNV or CNV) of a method of the invention is between 15 to 0.005%>, such as between 10 to 0.005%>, 10 to 0.01%>, 10 to 0.1 %>, 5 to 0.005%, 5 to 0.01%, 5 to 0.1%, 1 to 0.005%, 1 to 0.01%, 1 to 0.1%, 0.5 to 0.005%, 0.5 to 0.01%), 0.5 to 0.1%), or 0.1 to 0.01, inclusive. In some embodiments, a limit of detection is such that a mutation (such as an SNV or CNV) that is present in less than or equal to 10, 5, 2, 1, 0.5, 0.1, 0.05, 0.01, or 0.005% of the DNA or RNA molecules with that locus in a sample (such as a sample of cfDNA or cfRNA) is detected (or is capable of being detected). For example, the mutation can be detected even if less than or equal to 10, 5, 2, 1, 0.5, 0.1, 0.05, 0.01, or 0.005% of the DNA or RNA molecules that have that locus have that mutation in the locus (instead of, for example, a wild-type or non-mutated version of the locus or a different mutation at that locus). In some embodiments, a limit of detection is such that a mutation (such as an SNV or CNV) that is present in less than or equal to 10, 5, 2, 1, 0.5, 0.1, 0.05, 0.01, or 0.005% of the DNA or RNA molecules in a sample (such as a sample of cfDNA or cfRNA) is detected (or is capable of being detected). In some embodiments in which the CNV is a deletion, the deletion can be detected even if it is only present in less than or equal to 10, 5, 2, 1, 0.5, 0.1, 0.05, 0.01, or

0.005% of the DNA or RNA molecules that have a region of interest that may or may not contain the deletion in a sample. In some embodiments in which the CNV is a deletion, the deletion can be detected even if it is only present in less than or equal to 10, 5, 2, 1, 0.5, 0.1, 0.05, 0.01, or

0.005% of the DNA or RNA molecules in a sample. In some embodiments in which the CNV is a duplication, the duplication can be detected even if the extra duplicated DNA or RNA that is present is less than or equal to 10, 5, 2, 1, 0.5, 0.1, 0.05, 0.01, or 0.005% of the DNA or RNA molecules that have a region of interest that may or may not be duplicated in a sample in a sample. In some embodiments in

the likelihoods can still be combined.

**[302]** Exemplary methods of analyzing multiple factors or variables are disclosed in U.S. Patent No. 8,024,128 issued on September 20, 2011; U.S. Publication No. 2007/0027636, filed July 31, 2006; and U.S. Publication No. 2007/0178501, filed December 6, 2006, which are each hereby incorporated by reference in its entirety).

**[303]** In various embodiments, the combined probability of a particular hypothesis or diagnosis is greater than 80, 85, 90, 92, 94, 96, 98, 99, or 99.9%, or is greater than some other threshold value.

#### Limit of Detection

**[304]** In some embodiments, a limit of detection of a mutation (such as an SNV or CNV) of a method of the invention is less than or equal to 10, 5, 2, 1, 0.5, 0.1, 0.05, 0.01, or 0.005%. In some embodiments, a limit of detection of a mutation (such as an SNV or CNV) of a method of the invention is between 15 to 0.005%>, such as between 10 to 0.005%>, 10 to 0.01%>, 10 to 0.1 %>, 5 to 0.005%, 5 to 0.01%, 5 to 0.1%, 1 to 0.005%, 1 to 0.01%, 1 to 0.1%, 0.5 to 0.005%, 0.5 to 0.01%), 0.5 to 0.1%), or 0.1 to 0.01, inclusive. In some embodiments, a limit of detection is such that a mutation (such as an SNV or CNV) that is present in less than or equal to 10, 5, 2, 1, 0.5, 0.1, 0.05, 0.01, or 0.005% of the DNA or RNA molecules with that locus in a sample (such as a sample of cfDNA or cfRNA) is detected (or is capable of being detected). For example, the mutation can be detected even if less than or equal to 10, 5, 2, 1, 0.5, 0.1, 0.05, 0.01, or 0.005% of the DNA or RNA molecules that have that locus have that mutation in the locus (instead of, for example, a wild-type or non-mutated version of the locus or a different mutation at that locus). In some embodiments, a limit of detection is such that a mutation (such as an SNV or CNV) that is present in less than or equal to 10, 5, 2, 1, 0.5, 0.1, 0.05, 0.01, or 0.005% of the DNA or RNA molecules in a sample (such as a sample of cfDNA or cfRNA) is detected (or is capable of being detected). In some embodiments in which the CNV is a deletion, the deletion can be detected even if it is only present in less than or equal to 10, 5, 2, 1, 0.5, 0.1, 0.05, 0.01, or

0.005% of the DNA or RNA molecules that have a region of interest that may or may not contain the deletion in a sample. In some embodiments in which the CNV is a deletion, the deletion can be detected even if it is only present in less than or equal to 10, 5, 2, 1, 0.5, 0.1, 0.05, 0.01, or

0.005% of the DNA or RNA molecules in a sample. In some embodiments in which the CNV is a duplication, the duplication can be detected even if the extra duplicated DNA or RNA that is present is less than or equal to 10, 5, 2, 1, 0.5, 0.1, 0.05, 0.01, or 0.005% of the DNA or RNA molecules that have a region of interest that may or may not be duplicated in a sample in a sample. In some embodiments in

which the CNV is a duplication, the duplication can be detected even if the extra duplicated DNA or RNA that is present is less than or equal to 10, 5, 2,

1, 0.5, 0.1, 0.05, 0.01, or 0.005%) of the DNA or RNA molecules in a sample. Example 6 provides exemplary methods for calculating the limit of detection. In some embodiments, the “LOD-zs5.0-mr5” method of Example 6 is used.

### Exemplary Samples

**[305]** In some embodiments of any of the aspects of the invention, the sample includes cellular and/or extracellular genetic material from cells suspected of having a deletion or duplication, such as cells suspected of being cancerous. In some embodiments, the sample comprises any tissue or bodily fluid suspected of containing cells, DNA, or RNA having a deletion or duplication, such as cancer cells, DNA, or RNA. The genetic measurements used as part of these methods can be made on any sample comprising DNA or RNA, for example but not limited to, tissue, blood, serum, plasma, urine, hair, tears, saliva, skin, fingernails, feces, bile, lymph, cervical mucus, semen, or other cells or materials comprising nucleic acids. Samples may include any cell type or DNA or RNA from any cell type may be used (such as cells from any organ or tissue suspected of being cancerous, or neurons). In some embodiments, the sample includes nuclear and/or mitochondrial DNA. In some embodiments, the sample is from any of the target individuals disclosed herein. In some embodiments, the target individual is a born individual, a gestating fetus, a non-gestating fetus such as a products of conception sample, an embryo, or any other individual.

**[306]** Exemplary samples include those containing cfDNA or cfRNA. In some embodiments, cfDNA is available for analysis without requiring the step of lysing cells. Cell-free DNA may be obtained from a variety of tissues, such as tissues that are in liquid form, *e.g.*, blood, plasma, lymph, ascites fluid, or cerebral spinal fluid. In some cases, cfDNA is comprised of DNA derived from fetal cells. In some cases, cfDNA is comprised of DNA derived from both fetal and maternal cells. In some cases, the cfDNA is isolated from plasma that has been isolated from whole blood that has been centrifuged to remove cellular material. The cfDNA

may be a mixture of DNA derived from target cells (such as cancer cells) and non-target cells (such as non-cancer cells).

**[307]** In some embodiments, the sample contains or is suspected to contain a mixture of DNA (or RNA), such as mixture of cancer DNA (or RNA) and noncancerous DNA (or RNA). In some embodiments, at least 0.5, 1, 3, 5, 7, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 92, 94, 95, 96, 98, 99, or

which the CNV is a duplication, the duplication can be detected even if the extra duplicated DNA or RNA that is present is less than or equal to 10, 5, 2,

1, 0.5, 0.1, 0.05, 0.01, or 0.005%) of the DNA or RNA molecules in a sample. Example 6 provides exemplary methods for calculating the limit of detection. In some embodiments, the “LOD-zs5.0-mr5” method of Example 6 is used.

### Exemplary Samples

**[305]** In some embodiments of any of the aspects of the invention, the sample includes cellular and/or extracellular genetic material from cells suspected of having a deletion or duplication, such as cells suspected of being cancerous. In some embodiments, the sample comprises any tissue or bodily fluid suspected of containing cells, DNA, or RNA having a deletion or duplication, such as cancer cells, DNA, or RNA. The genetic measurements used as part of these methods can be made on any sample comprising DNA or RNA, for example but not limited to, tissue, blood, serum, plasma, urine, hair, tears, saliva, skin, fingernails, feces, bile, lymph, cervical mucus, semen, or other cells or materials comprising nucleic acids. Samples may include any cell type or DNA or RNA from any cell type may be used (such as cells from any organ or tissue suspected of being cancerous, or neurons). In some embodiments, the sample includes nuclear and/or mitochondrial DNA. In some embodiments, the sample is from any of the target individuals disclosed herein. In some embodiments, the target individual is a born individual, a gestating fetus, a non-gestating fetus such as a products of conception sample, an embryo, or any other individual.

**[306]** Exemplary samples include those containing cfDNA or cfRNA. In some embodiments, cfDNA is available for analysis without requiring the step of lysing cells. Cell-free DNA may be obtained from a variety of tissues, such as tissues that are in liquid form, *e.g.*, blood, plasma, lymph, ascites fluid, or cerebral spinal fluid. In some cases, cfDNA is comprised of DNA derived from fetal cells. In some cases, cfDNA is comprised of DNA derived from both fetal and maternal cells. In some cases, the cfDNA is isolated from plasma that has been isolated from whole blood that has been centrifuged to remove cellular material. The cfDNA

may be a mixture of DNA derived from target cells (such as cancer cells) and non-target cells (such as non-cancer cells).

**[307]** In some embodiments, the sample contains or is suspected to contain a mixture of DNA (or RNA), such as mixture of cancer DNA (or RNA) and noncancerous DNA (or RNA). In some embodiments, at least 0.5, 1, 3, 5, 7, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 92, 94, 95, 96, 98, 99, or

100% of the cells in the sample are cancer cells. In some embodiments, at least 0.5,

1, 3, 5, 7, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 92, 94, 95, 96, 98, 99, or 100% of the DNA (such as cfDNA) or RNA (such as cfRNA) in the sample is from cancer cell(s). In various embodiments, the percent of cells in the sample that are cancerous cells is between 0.5 to 99%, such as between 1 to 95%, 5 to 95%, 10 to 90%, 5 to 70%, 10 to 70%, 20 to 90%, or 20 to 70%, inclusive. In some embodiments, the sample is enriched for cancer cells or for DNA or RNA from cancer cells. In some embodiments in which the sample is enriched for cancer cells, at least 0.5, 1, 2, 3, 4, 5, 6, 7, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 92, 94, 95, 96, 98, 99, or 100% of the cells in the enriched sample are cancer cells. In some embodiments in which the sample is enriched for DNA or RNA from cancer cells, at least 0.5, 1, 2, 3, 4, 5, 6, 7, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 92, 94, 95, 96, 98, 99, or 100% of the DNA or RNA in the enriched sample is from cancer cell(s). In some embodiments, cell sorting (such as Fluorescent Activated Cell Sorting (FACS)) is used to enrich for cancer cells (Barteneva *et al.*, Biochim Biophys Acta., 1836(1):105-22, Aug 2013. doi: 10.1016/j.bbcan.2013.02.004. Epub 2013 Feb 24, and Ibrahim *et al.*, Adv Biochem Eng Biotechnol. 106:19-39, 2007, which are each hereby incorporated by reference in its entirety).

**[308]** In some embodiments of any of the aspects of the invention, the sample comprises any tissue suspected of being at least partially of fetal origin. In some embodiments, the sample includes cellular and/or extracellular genetic material from the fetus, contaminating cellular and/or extracellular genetic material (such as genetic material from the mother of the fetus), or combinations thereof. In some embodiments, the sample comprises cellular genetic material from the fetus, contaminating cellular genetic material, or combinations thereof.

**[309]** In some embodiments, the sample is from a gestating fetus. In some embodiments, the sample is from a non-gestating fetus, such as a products of conception sample or a sample from any fetal tissue after fetal demise. In some embodiments, the sample is a maternal whole blood sample, cells isolated from a maternal blood sample, maternal plasma sample, maternal serum sample, amniocentesis sample, placental tissue sample (*e.g.*, chorionic villus, decidua, or placental membrane), cervical mucus sample, or other sample from a fetus. In some embodiments, at least 3, 5, 7, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 92, 94, 95, 96, 98, 99, or 100% of the cells in the sample are maternal cells. In various embodiments, the percent of cells in the sample that are maternal cells is between 5 to 99%, such as between 10 to 95%, 20 to **95%**, 30 to 90%<sub>o</sub>, 30 to 70%<sub>></sub>, 40 to 90%<sub>></sub>, 40 to 70%<sub>></sub>, 50 to 90%<sub>></sub>, or

100% of the cells in the sample are cancer cells. In some embodiments, at least 0.5, 1, 3, 5, 7, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 92, 94, 95, 96, 98, 99, or 100% of the DNA (such as cfDNA) or RNA (such as cfRNA) in the sample is from cancer cell(s). In various embodiments, the percent of cells in the sample that are cancerous cells is between 0.5 to 99%, such as between 1 to 95%, 5 to 95%, 10 to 90%, 5 to 70%, 10 to 70%, 20 to 90%, or 20 to 70%, inclusive. In some embodiments, the sample is enriched for cancer cells or for DNA or RNA from cancer cells. In some embodiments in which the sample is enriched for cancer cells, at least 0.5, 1, 2, 3, 4, 5, 6, 7, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 92, 94, 95, 96, 98, 99, or 100% of the cells in the enriched sample are cancer cells. In some embodiments in which the sample is enriched for DNA or RNA from cancer cells, at least 0.5, 1, 2, 3, 4, 5, 6, 7, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 92, 94, 95, 96, 98, 99, or 100% of the DNA or RNA in the enriched sample is from cancer cell(s). In some embodiments, cell sorting (such as Fluorescent Activated Cell Sorting (FACS)) is used to enrich for cancer cells (Barteneva *et al.*, Biochim Biophys Acta., 1836(1):105-22, Aug 2013. doi: 10.1016/j.bbcan.2013.02.004. Epub 2013 Feb 24, and Ibrahim *et al.*, Adv Biochem Eng Biotechnol. 106:19-39, 2007, which are each hereby incorporated by reference in its entirety).

**[308]** In some embodiments of any of the aspects of the invention, the sample comprises any tissue suspected of being at least partially of fetal origin. In some embodiments, the sample includes cellular and/or extracellular genetic material from the fetus, contaminating cellular and/or extracellular genetic material (such as genetic material from the mother of the fetus), or combinations thereof. In some embodiments, the sample comprises cellular genetic material from the fetus, contaminating cellular genetic material, or combinations thereof.

**[309]** In some embodiments, the sample is from a gestating fetus. In some embodiments, the sample is from a non-gestating fetus, such as a products of conception sample or a sample from any fetal tissue after fetal demise. In some embodiments, the sample is a maternal whole blood sample, cells isolated from a maternal blood sample, maternal plasma sample, maternal serum sample, amniocentesis sample, placental tissue sample (*e.g.*, chorionic villus, decidua, or placental membrane), cervical mucus sample, or other sample from a fetus. In some embodiments, at least 3, 5, 7, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 92, 94, 95, 96, 98, 99, or 100% of the cells in the sample are maternal cells. In various embodiments, the percent of cells in the sample that are maternal cells is between 5 to 99%, such as between 10 to 95%, 20 to 95%, 30 to 90%, 30 to 70%, 40 to 90%, 40 to 70%, 50 to 90%, or

50 to 80%>, inclusive.

**[310]** In some embodiments, the sample is enriched for fetal cells. In some embodiments in which the sample is enriched for fetal cells, at least 0.5, 1, 2, 3, 4, 5, 6, 7% or more of the cells in the enriched sample are fetal cells. In some embodiments, the percent of cells in the sample that are fetal cells is between 0.5 to 100%>, such as between 1 to 99%, 5 to 95%, 10 to 95%, 10 to 95%), 20 to 90%), or 30 to 70%, inclusive. In some embodiments, the sample is enriched for fetal DNA. In some embodiments in which the sample is enriched for fetal DNA, at least 0.5, 1,

2, 3, 4, 5, 6, 7% or more of the DNA in the enriched sample is fetal DNA. In some embodiments, the percent of DNA in the sample that is fetal DNA is between 0.5 to 100%, such as between 1 to 99%, 5 to 95%, 10 to 95%, 10 to 95%, 20 to 90%, or 30 to 70%, inclusive.

**[311]** In some embodiments, the sample includes a single cell or includes DNA and/or RNA from a single cell. In some embodiments, multiple individual cells (*e.g.*, at least 5, 10, 20, 30, 40, or 50 cells from the same subject or from different subjects) are analyzed in parallel. In some embodiments, cells from multiple samples from the same individual are combined, which reduces the amount of work compared to analyzing the samples separately. Combining multiple samples can also allow multiple tissues to be tested for cancer simultaneously (which can be used to provide or more thorough screening for cancer or to determine whether cancer may have metastasized to other tissues).

**[312]** In some embodiments, the sample contains a single cell or a small number of cells, such as 2, 3, 5, 6, 7, 8, 9, or 10 cells. In some embodiments, the sample has between 1 to 100, 100 to 500, or 500 to 1,000 cells, inclusive. In some embodiments, the sample contains 1 to 10 picograms, 10 to 100 picograms, 100 picograms to 1 nanogram, 1 to 10 nanograms, 10 to 100 nanograms, or 100 nanograms to 1 microgram of RNA and/or DNA, inclusive.

**[313]** In some embodiments, the sample is embedded in parafilm. In some embodiments, the sample is preserved with a preservative such as formaldehyde and optionally encased in paraffin, which may cause cross-linking of the DNA such that less of it is available for PCR. In some embodiments, the sample is a formaldehyde fixed-paraffin embedded (FFPE) sample. In some embodiments, the sample is a fresh sample (such as a sample obtained with 1 or 2 days of analysis). In some embodiments, the sample is frozen prior to analysis. In some embodiments, the sample is a historical sample.

**[314]** These samples can be used in any of the methods of the invention.

#### Exemplary Sample Analysis Methods

## CLAIMS

What is claimed is:

1. A method for determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of one or more cells from an individual comprising:

obtaining phased genetic data for the first homologous chromosome segment comprising, for each locus in a set of polymorphic loci on the first homologous chromosome segment, the identity of the allele present at that locus on the first homologous chromosome segment,

obtaining phased genetic data for the second homologous chromosome segment comprising, for each locus in the set of polymorphic loci on the second homologous chromosome segment, the identity of the allele present at that locus on the second homologous chromosome segment,

obtaining measured genetic allelic data comprising, for each of the alleles at each of the loci in the set of polymorphic loci, the amount of each allele present in a sample of DNA from one or more target cells and one or more non-target cells from the individual;

enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment;

calculating, for each of the hypotheses, expected genetic data for the plurality of loci in the sample from the obtained phased genetic data for one or more possible ratios of DNA from the one or more target cells to the total DNA in the sample;

calculating, on a computer, for each possible ratio of DNA and for each hypothesis, the data fit between the obtained genetic data of the sample and the expected genetic data for the sample for that possible ratio of DNA and for that hypothesis;

ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more cells from the individual.

2. A method for determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of one or more cells from an individual comprising:

obtaining phased genetic data for the first homologous chromosome segment comprising, for

## CLAIMS

What is claimed is:

1. A method for determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of one or more cells from an individual comprising:

obtaining phased genetic data for the first homologous chromosome segment comprising, for each locus in a set of polymorphic loci on the first homologous chromosome segment, the identity of the allele present at that locus on the first homologous chromosome segment,

obtaining phased genetic data for the second homologous chromosome segment comprising, for each locus in the set of polymorphic loci on the second homologous chromosome segment, the identity of the allele present at that locus on the second homologous chromosome segment,

obtaining measured genetic allelic data comprising, for each of the alleles at each of the loci in the set of polymorphic loci, the amount of each allele present in a sample of DNA from one or more target cells and one or more non-target cells from the individual;

enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment;

calculating, for each of the hypotheses, expected genetic data for the plurality of loci in the sample from the obtained phased genetic data for one or more possible ratios of DNA from the one or more target cells to the total DNA in the sample;

calculating, on a computer, for each possible ratio of DNA and for each hypothesis, the data fit between the obtained genetic data of the sample and the expected genetic data for the sample for that possible ratio of DNA and for that hypothesis;

ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more cells from the individual.

2. A method for determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of one or more cells from an individual comprising:

obtaining phased genetic data for the first homologous chromosome segment comprising, for

each locus in a set of polymorphic loci on the first homologous chromosome segment, the identity of the allele present at that locus on the first homologous chromosome segment,

obtaining phased genetic data for the second homologous chromosome segment comprising, for each locus in the set of polymorphic loci on the second homologous chromosome segment, the identity of the allele present at that locus on the second homologous chromosome segment,

obtaining measured genetic allelic data comprising, for each of the alleles at each of the loci in the set of polymorphic loci, the amount of each allele present in a sample of DNA from one or more target cells and one or more non-target cells from the individual;

enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment;

calculating, for each of the hypotheses, expected genetic data for the plurality of loci in the sample from the obtained phased genetic data for one or more possible ratios of DNA from the one or more target cells to the total DNA in the sample;

calculating, on a computer, for each locus in the plurality of loci, each possible ratio of DNA, and each hypothesis, the likelihood that the hypothesis is correct by comparing the obtained genetic data of the sample for that locus and the expected genetic data for that locus for that possible ratio of DNA and for that hypothesis;

determining the combined probability for each hypothesis by combining the probabilities of that hypothesis for each locus and each possible ratio; and

selecting the hypothesis with the greatest combined probability, thereby determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment.

3. A method for determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of one or more cells from an individual comprising:

obtaining phased genetic data for the first homologous chromosome segment comprising, for each locus in a set of polymorphic loci on the first homologous chromosome segment, the identity of the allele present at that locus on the first homologous chromosome segment,

obtaining phased genetic data for the second homologous chromosome segment comprising, for each locus in the set of polymorphic loci on the second homologous chromosome segment, the identity of the allele present at that locus on the second homologous chromosome segment,

each locus in a set of polymorphic loci on the first homologous chromosome segment, the identity of the allele present at that locus on the first homologous chromosome segment,

obtaining phased genetic data for the second homologous chromosome segment comprising, for each locus in the set of polymorphic loci on the second homologous chromosome segment, the identity of the allele present at that locus on the second homologous chromosome segment,

obtaining measured genetic allelic data comprising, for each of the alleles at each of the loci in the set of polymorphic loci, the amount of each allele present in a sample of DNA from one or more target cells and one or more non-target cells from the individual;

enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment;

calculating, for each of the hypotheses, expected genetic data for the plurality of loci in the sample from the obtained phased genetic data for one or more possible ratios of DNA from the one or more target cells to the total DNA in the sample;

calculating, on a computer, for each locus in the plurality of loci, each possible ratio of DNA, and each hypothesis, the likelihood that the hypothesis is correct by comparing the obtained genetic data of the sample for that locus and the expected genetic data for that locus for that possible ratio of DNA and for that hypothesis;

determining the combined probability for each hypothesis by combining the probabilities of that hypothesis for each locus and each possible ratio; and

selecting the hypothesis with the greatest combined probability, thereby determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment.

3. A method for determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in the genome of one or more cells from an individual comprising:

obtaining phased genetic data for the first homologous chromosome segment comprising, for each locus in a set of polymorphic loci on the first homologous chromosome segment, the identity of the allele present at that locus on the first homologous chromosome segment,

obtaining phased genetic data for the second homologous chromosome segment comprising, for each locus in the set of polymorphic loci on the second homologous chromosome segment, the identity of the allele present at that locus on the second homologous chromosome segment,

obtaining measured genetic allelic data comprising, for each of the alleles at each of the loci in the set of polymorphic loci, the amount of each allele present in a sample of DNA from one or more cells from the individual;

calculating allele ratios for one or more loci in the set of polymorphic loci that are heterozygous in at least one cell; wherein the calculated allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus; and

determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment by comparing, for one or more of the calculated allele ratios, the calculated allele ratio for a locus to an allele ratio that is expected for that locus if the first and second homologous chromosome segments are present in equal proportions.

4. The method of claim 3, wherein a calculated allele ratio is indicative of an overrepresentation of the number of copies of the first homologous chromosome segment if either:

(i) the allele ratio for the measured quantity of the allele present at that locus on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is greater than the expected allele ratio for that locus, or

(ii) the allele ratio for the measured quantity of the allele present at that locus on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than the expected allele ratio for that locus.

5. The method of claim 3, wherein a calculated allele ratio is indicative of no overrepresentation of the number of copies of the first homologous chromosome segment if either:

(i) the allele ratio for the measured quantity of the allele present at that locus on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than or equal to the expected allele ratio for that locus, or

(ii) the allele ratio for the measured quantity of the allele present at that locus on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is greater than or equal to the expected allele ratio for that locus.

6. A method for determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in cell-free DNA from an individual comprising:

obtaining measured genetic allelic data comprising, for each of the alleles at each of the loci in the set of polymorphic loci, the amount of each allele present in a sample of DNA from one or more cells from the individual;

calculating allele ratios for one or more loci in the set of polymorphic loci that are heterozygous in at least one cell; wherein the calculated allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus; and

determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment by comparing, for one or more of the calculated allele ratios, the calculated allele ratio for a locus to an allele ratio that is expected for that locus if the first and second homologous chromosome segments are present in equal proportions.

4. The method of claim 3, wherein a calculated allele ratio is indicative of an overrepresentation of the number of copies of the first homologous chromosome segment if either:

(i) the allele ratio for the measured quantity of the allele present at that locus on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is greater than the expected allele ratio for that locus, or

(ii) the allele ratio for the measured quantity of the allele present at that locus on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than the expected allele ratio for that locus.

5. The method of claim 3, wherein a calculated allele ratio is indicative of no overrepresentation of the number of copies of the first homologous chromosome segment if either:

(i) the allele ratio for the measured quantity of the allele present at that locus on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than or equal to the expected allele ratio for that locus, or

(ii) the allele ratio for the measured quantity of the allele present at that locus on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is greater than or equal to the expected allele ratio for that locus.

6. A method for determining if there is an overrepresentation of the number of copies of a first homologous chromosome segment as compared to a second homologous chromosome segment in cell-free DNA from an individual comprising:

obtaining phased genetic data for the first homologous chromosome segment comprising, for each locus in a set of polymorphic loci on the first homologous chromosome segment, the identity of the allele present at that locus on the first homologous chromosome segment,

obtaining phased genetic data for the second homologous chromosome segment comprising, for each locus in the set of polymorphic loci on the second homologous chromosome segment, the identity of the allele present at that locus on the second homologous chromosome segment,

obtaining measured genetic allelic data comprising, for each of the alleles at each of the loci in the set of polymorphic loci, the amount of each allele present in a mixed sample of cell-free DNA from the individual; wherein the mixed sample comprises cell-free DNA from two or more genetically different cells;

calculating allele ratios for one or more loci in the set of polymorphic loci that are heterozygous in at least one cell; wherein the calculated allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus; and

determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment by comparing, for one or more of the calculated allele ratios, the calculated allele ratio for a locus to an allele ratio that is expected for that locus if the first and second homologous chromosome segments are present in equal proportions;

wherein a calculated allele ratio is indicative of an overrepresentation of the number of copies of the first homologous chromosome segment if either:

- (i) the allele ratio for the measured quantity of the allele present at that locus on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is greater than the expected allele ratio for that locus, or
- (ii) the allele ratio for the measured quantity of the allele present at that locus on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than the expected allele ratio for that locus; or

wherein a calculated allele ratio is indicative of no overrepresentation of the number of copies of the first homologous chromosome segment if either:

- (i) the allele ratio for the measured quantity of the allele present at that locus on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than or equal to the expected allele ratio for that locus, or

obtaining phased genetic data for the first homologous chromosome segment comprising, for each locus in a set of polymorphic loci on the first homologous chromosome segment, the identity of the allele present at that locus on the first homologous chromosome segment,

obtaining phased genetic data for the second homologous chromosome segment comprising, for each locus in the set of polymorphic loci on the second homologous chromosome segment, the identity of the allele present at that locus on the second homologous chromosome segment,

obtaining measured genetic allelic data comprising, for each of the alleles at each of the loci in the set of polymorphic loci, the amount of each allele present in a mixed sample of cell-free DNA from the individual; wherein the mixed sample comprises cell-free DNA from two or more genetically different cells;

calculating allele ratios for one or more loci in the set of polymorphic loci that are heterozygous in at least one cell; wherein the calculated allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus; and

determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment by comparing, for one or more of the calculated allele ratios, the calculated allele ratio for a locus to an allele ratio that is expected for that locus if the first and second homologous chromosome segments are present in equal proportions;

wherein a calculated allele ratio is indicative of an overrepresentation of the number of copies of the first homologous chromosome segment if either:

- (i) the allele ratio for the measured quantity of the allele present at that locus on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is greater than the expected allele ratio for that locus, or
- (ii) the allele ratio for the measured quantity of the allele present at that locus on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than the expected allele ratio for that locus; or

wherein a calculated allele ratio is indicative of no overrepresentation of the number of copies of the first homologous chromosome segment if either:

- (i) the allele ratio for the measured quantity of the allele present at that locus on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than or equal to the expected allele ratio for that locus, or

- (ii) the allele ratio for the measured quantity of the allele present at that locus on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is greater than or equal to the expected allele ratio for that locus.

7. The method of any one of claims 3-6, wherein each calculated allele ratio is the allele ratio for the measured quantity of the allele present at that locus on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus, wherein (i) a calculated allele ratio is indicative of an overrepresentation of the number of copies of the first homologous chromosome segment if the calculated allele ratio is greater than the expected allele ratio for that locus, or (ii) a calculated allele ratio is indicative of no overrepresentation of the number of copies of the first homologous chromosome segment if the calculated ratio is less than or equal to the expected allele ratio for that locus.

8. The method of any one of claims 3-7, comprising  
calculating allele ratios for a plurality of loci in the set of polymorphic loci that are heterozygous in at least one cell;

determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment by comparing the calculated allele ratios to the expected allele ratios.

9. The method of any one of claims 3-8, wherein comparing the calculated allele ratios to the expected allele ratios comprises:

- (i) determining, for each calculated allele ratio, whether the calculated allele ratio is above or below the expected allele ratio for a particular locus irrespective of the magnitude of the difference;

- (ii) determining, for each calculated allele ratio, the magnitude of the difference between the calculated allele ratio and the expected allele ratio for a particular locus irrespective of whether the calculated allele ratio is above or below the expected allele ratio;

- (iii) determining, for each calculated allele ratio, whether the calculated allele ratio is above or below the expected allele ratio and the magnitude of the difference for a particular locus;

- (iv) determining whether an average or weighted average value for the calculated allele ratios is above or below an average or weighted average value of the expected allele ratios irrespective of the magnitude of the difference;

- (v) determining the magnitude of the difference between an average or weighted average value for the calculated allele ratios and an average or weighted average value for the expected allele ratios

- (ii) the allele ratio for the measured quantity of the allele present at that locus on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is greater than or equal to the expected allele ratio for that locus.

7. The method of any one of claims 3-6, wherein each calculated allele ratio is the allele ratio for the measured quantity of the allele present at that locus on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus, wherein (i) a calculated allele ratio is indicative of an overrepresentation of the number of copies of the first homologous chromosome segment if the calculated allele ratio is greater than the expected allele ratio for that locus, or (ii) a calculated allele ratio is indicative of no overrepresentation of the number of copies of the first homologous chromosome segment if the calculated ratio is less than or equal to the expected allele ratio for that locus.

8. The method of any one of claims 3-7, comprising  
calculating allele ratios for a plurality of loci in the set of polymorphic loci that are heterozygous in at least one cell;

determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment by comparing the calculated allele ratios to the expected allele ratios.

9. The method of any one of claims 3-8, wherein comparing the calculated allele ratios to the expected allele ratios comprises:

- (i) determining, for each calculated allele ratio, whether the calculated allele ratio is above or below the expected allele ratio for a particular locus irrespective of the magnitude of the difference;

- (ii) determining, for each calculated allele ratio, the magnitude of the difference between the calculated allele ratio and the expected allele ratio for a particular locus irrespective of whether the calculated allele ratio is above or below the expected allele ratio;

- (iii) determining, for each calculated allele ratio, whether the calculated allele ratio is above or below the expected allele ratio and the magnitude of the difference for a particular locus;

- (iv) determining whether an average or weighted average value for the calculated allele ratios is above or below an average or weighted average value of the expected allele ratios irrespective of the magnitude of the difference;

- (v) determining the magnitude of the difference between an average or weighted average value for the calculated allele ratios and an average or weighted average value for the expected allele ratios

irrespective of whether the average value or weighted average value for the calculated allele ratio is above or below the average or weighted average value of the expected allele ratio;

(vi) determining whether an average or weighted average value for the calculated allele ratios is above or below an average or weighted average value for the expected allele ratios and the magnitude of the difference;

(vii) determining an average or weighted average value of the magnitude of the difference between the calculated allele ratios and the expected allele ratios; or

(viii) a combination of two or more of (i) to (vii).

10. The method of any one of claims 3-9, comprising determining that there is an overrepresentation of the number of copies of the first homologous chromosome segment if:

(i) the number of calculated allele ratios that are indicative of an overrepresentation of the number of copies of the first homologous chromosome segment is above a first threshold value,

(ii) the number of calculated allele ratios that are indicative of no overrepresentation of the number of copies of the first homologous chromosome segment is below a second threshold value;

(iii) the magnitude of the difference between the calculated allele ratios that are indicative of an overrepresentation of the number of copies of the first homologous chromosome segment and the corresponding expected allele ratios is above a third threshold value,

(iv) the magnitude of the difference between the calculated allele ratios that are indicative of no overrepresentation of the number of copies of the first homologous chromosome segment and the corresponding expected allele ratios is below a fourth threshold value,

(v) the average or weighted average value of the calculated allele ratios for the measured quantity of the allele present on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is greater than the average or weighted average value of the expected allele ratios by at least a fifth threshold value,

(vi) the average or weighted average value of the calculated allele ratios for the measured quantity of the allele present on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than the average or weighted average value of the expected allele ratios by at least a sixth threshold value,

(vii) the data fit between the calculated allele ratios and allele ratios that are predicted for an overrepresentation of the number of copies of the first homologous chromosome segment is below a

irrespective of whether the average value or weighted average value for the calculated allele ratio is above or below the average or weighted average value of the expected allele ratio;

(vi) determining whether an average or weighted average value for the calculated allele ratios is above or below an average or weighted average value for the expected allele ratios and the magnitude of the difference;

(vii) determining an average or weighted average value of the magnitude of the difference between the calculated allele ratios and the expected allele ratios; or

(viii) a combination of two or more of (i) to (vii).

10. The method of any one of claims 3-9, comprising determining that there is an overrepresentation of the number of copies of the first homologous chromosome segment if:

(i) the number of calculated allele ratios that are indicative of an overrepresentation of the number of copies of the first homologous chromosome segment is above a first threshold value,

(ii) the number of calculated allele ratios that are indicative of no overrepresentation of the number of copies of the first homologous chromosome segment is below a second threshold value;

(iii) the magnitude of the difference between the calculated allele ratios that are indicative of an overrepresentation of the number of copies of the first homologous chromosome segment and the corresponding expected allele ratios is above a third threshold value,

(iv) the magnitude of the difference between the calculated allele ratios that are indicative of no overrepresentation of the number of copies of the first homologous chromosome segment and the corresponding expected allele ratios is below a fourth threshold value,

(v) the average or weighted average value of the calculated allele ratios for the measured quantity of the allele present on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus is greater than the average or weighted average value of the expected allele ratios by at least a fifth threshold value,

(vi) the average or weighted average value of the calculated allele ratios for the measured quantity of the allele present on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than the average or weighted average value of the expected allele ratios by at least a sixth threshold value,

(vii) the data fit between the calculated allele ratios and allele ratios that are predicted for an overrepresentation of the number of copies of the first homologous chromosome segment is below a

seventh threshold value;

(viii) the data fit between the calculated allele ratios and allele ratios that are predicted for no overrepresentation of the number of copies of the first homologous chromosome segment is above an eight threshold value; or

(ix) a combination of two or more of (i) to (viii).

11. The method of any one of claims 3-10, comprising determining that there is not an overrepresentation of the number of copies of the first homologous chromosome segment if:

(i) the number of calculated allele ratios that are indicative of an overrepresentation of the number of copies of the first homologous chromosome segment is below the first threshold value,

(ii) the number of calculated allele ratios that are indicative of no overrepresentation of the number of copies of the first homologous chromosome segment is above the second threshold value,

(iii) the magnitude of the difference between the calculated allele ratios that are indicative of an overrepresentation of the number of copies of the first homologous chromosome segment and the corresponding expected allele ratios is below a third threshold value,

(iv) the magnitude of the difference between the calculated allele ratios that are indicative of no overrepresentation of the number of copies of the first homologous chromosome segment and the corresponding expected allele ratios is above a fourth threshold value,

(v) the average or weighted average value of the calculated allele ratios for the measured quantity of the allele present on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus minus the average or weighted average value of the expected allele ratios is less than a fifth threshold value,

(vi) the average or weighted average value of the expected allele ratios minus the average or weighted average value of the calculated allele ratios for the measured quantity of the allele present on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than a sixth threshold value,

(vii) the data fit between the calculated allele ratios and allele ratios that are predicted for an overrepresentation of the number of copies of the first homologous chromosome segment is above a seventh threshold value;

(viii) the data fit between the calculated allele ratios and allele ratios that are predicted for no overrepresentation of the number of copies of the first homologous chromosome segment is below an eight threshold value; or

seventh threshold value;

(viii) the data fit between the calculated allele ratios and allele ratios that are predicted for no overrepresentation of the number of copies of the first homologous chromosome segment is above an eight threshold value; or

(ix) a combination of two or more of (i) to (viii).

11. The method of any one of claims 3-10, comprising determining that there is not an overrepresentation of the number of copies of the first homologous chromosome segment if:

(i) the number of calculated allele ratios that are indicative of an overrepresentation of the number of copies of the first homologous chromosome segment is below the first threshold value,

(ii) the number of calculated allele ratios that are indicative of no overrepresentation of the number of copies of the first homologous chromosome segment is above the second threshold value,

(iii) the magnitude of the difference between the calculated allele ratios that are indicative of an overrepresentation of the number of copies of the first homologous chromosome segment and the corresponding expected allele ratios is below a third threshold value,

(iv) the magnitude of the difference between the calculated allele ratios that are indicative of no overrepresentation of the number of copies of the first homologous chromosome segment and the corresponding expected allele ratios is above a fourth threshold value,

(v) the average or weighted average value of the calculated allele ratios for the measured quantity of the allele present on the first homologous chromosome divided by the total measured quantity of all the alleles for the locus minus the average or weighted average value of the expected allele ratios is less than a fifth threshold value,

(vi) the average or weighted average value of the expected allele ratios minus the average or weighted average value of the calculated allele ratios for the measured quantity of the allele present on the second homologous chromosome divided by the total measured quantity of all the alleles for the locus is less than a sixth threshold value,

(vii) the data fit between the calculated allele ratios and allele ratios that are predicted for an overrepresentation of the number of copies of the first homologous chromosome segment is above a seventh threshold value;

(viii) the data fit between the calculated allele ratios and allele ratios that are predicted for no overrepresentation of the number of copies of the first homologous chromosome segment is below an eight threshold value; or

(ix) a combination of two or more of (i) to (viii).

12. The method of any one of claims 3-11, wherein determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment further comprises:

enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment;

estimating, for each hypothesis, either (i) predicted allele ratios for the loci that are heterozygous in at least one cell given the degree of overrepresentation specified by that hypothesis or (ii) an expected distribution of a test statistic calculated using the predicted allele ratios,

calculating, for each hypothesis, the likelihood that the hypothesis is correct by comparing either (i) the calculated allele ratios to the predicted allele ratios, or (ii) a test statistic calculated using the calculated allele ratios to the expected distribution of the test statistic calculated using the predicted allele ratios; and

determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment by selecting the hypothesis with the greatest likelihood.

13. The method of any one of claims 1-12, wherein the phased genetic data comprises probabilistic data.

14. The method of any one of claims 3-13, wherein determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment further comprises:

enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment;

estimating, for each hypothesis, predicted allele ratios for the loci that are heterozygous in at least one cell given the phased genetic data for the first homologous chromosome segment, the phased genetic data for the second homologous chromosome segment, and the degree of overrepresentation specified by that hypothesis,

calculating, for each hypothesis, the likelihood that the hypothesis is correct by comparing the calculated allele ratios to the predicted allele ratios; and

determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment by selecting the hypothesis with the greatest likelihood.

(ix) a combination of two or more of (i) to (viii).

12. The method of any one of claims 3-11, wherein determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment further comprises:

enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment;

estimating, for each hypothesis, either (i) predicted allele ratios for the loci that are heterozygous in at least one cell given the degree of overrepresentation specified by that hypothesis or (ii) an expected distribution of a test statistic calculated using the predicted allele ratios,

calculating, for each hypothesis, the likelihood that the hypothesis is correct by comparing either (i) the calculated allele ratios to the predicted allele ratios, or (ii) a test statistic calculated using the calculated allele ratios to the expected distribution of the test statistic calculated using the predicted allele ratios; and

determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment by selecting the hypothesis with the greatest likelihood.

13. The method of any one of claims 1-12, wherein the phased genetic data comprises probabilistic data.

14. The method of any one of claims 3-13, wherein determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment further comprises:

enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment;

estimating, for each hypothesis, predicted allele ratios for the loci that are heterozygous in at least one cell given the phased genetic data for the first homologous chromosome segment, the phased genetic data for the second homologous chromosome segment, and the degree of overrepresentation specified by that hypothesis,

calculating, for each hypothesis, the likelihood that the hypothesis is correct by comparing the calculated allele ratios to the predicted allele ratios; and

determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment by selecting the hypothesis with the greatest likelihood.

15. The method of any one of claims 1-14, wherein the overrepresentation of the number of copies of the first homologous chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment in the genome of one or more of the cells.

16. The method of claim 15, wherein the total measured quantity of all the alleles for one or more of the loci is compared to a reference amount to determine whether the overrepresentation of the number of copies of the first homologous chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment in the genome of one or more of the cells.

17. The method of claim 15, wherein the magnitude of the difference between the calculated allele ratio and the expected allele ratio for one or more loci is used to determine whether the overrepresentation of the number of copies of the first homologous chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment in the genome of one or more of the cells.

18. The method of any one of claims 3-17, wherein the sample comprises DNA from a mixture of one or more target cells and one or more non-target cells, and wherein the method further comprises:

calculating the ratio of DNA from the one or more target cells to the total DNA in the sample;

enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment;

estimating, for each hypothesis, either (i) predicted allele ratios for the loci that are heterozygous in at least one cell given calculated ratio of DNA and the degree of overrepresentation specified by that hypothesis or (ii) an expected distribution of a test statistic calculated using the predicted allele ratios and the calculated ratio of DNA,

calculating, for each hypothesis, the likelihood that the hypothesis is correct by comparing either (i) the calculated allele ratios to the predicted allele ratios, or (ii) a test statistic calculated using the calculated allele ratios and the calculated ratio of DNA to the expected distribution of the test statistic calculated using the predicted allele ratios and the calculated ratio of DNA; and

determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment by selecting the hypothesis with the greatest likelihood.

15. The method of any one of claims 1-14, wherein the overrepresentation of the number of copies of the first homologous chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment in the genome of one or more of the cells.

16. The method of claim 15, wherein the total measured quantity of all the alleles for one or more of the loci is compared to a reference amount to determine whether the overrepresentation of the number of copies of the first homologous chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment in the genome of one or more of the cells.

17. The method of claim 15, wherein the magnitude of the difference between the calculated allele ratio and the expected allele ratio for one or more loci is used to determine whether the overrepresentation of the number of copies of the first homologous chromosome segment is due to a duplication of the first homologous chromosome segment or a deletion of the second homologous chromosome segment in the genome of one or more of the cells.

18. The method of any one of claims 3-17, wherein the sample comprises DNA from a mixture of one or more target cells and one or more non-target cells, and wherein the method further comprises:

calculating the ratio of DNA from the one or more target cells to the total DNA in the sample;

enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment;

estimating, for each hypothesis, either (i) predicted allele ratios for the loci that are heterozygous in at least one cell given calculated ratio of DNA and the degree of overrepresentation specified by that hypothesis or (ii) an expected distribution of a test statistic calculated using the predicted allele ratios and the calculated ratio of DNA,

calculating, for each hypothesis, the likelihood that the hypothesis is correct by comparing either (i) the calculated allele ratios to the predicted allele ratios, or (ii) a test statistic calculated using the calculated allele ratios and the calculated ratio of DNA to the expected distribution of the test statistic calculated using the predicted allele ratios and the calculated ratio of DNA; and

determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment by selecting the hypothesis with the greatest likelihood.

19. The method of any one of claims 3-18, wherein the sample comprises DNA from a mixture of one or more target cells and one or more non-target cells, and wherein the method further comprises:

enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment;

estimating, for each hypothesis, either (i) predicted allele ratios for the loci that are heterozygous in at least one cell given the degree of overrepresentation specified by that hypothesis or (ii) for one or more possible ratios of DNA, an expected distribution of a test statistic calculated using the predicted allele ratios and the possible ratio of DNA from the one or more target cells to the total DNA in the sample;

calculating, for each possible ratio of DNA, a data fit by comparing either (i) the calculated allele ratios to the predicted allele ratios, or (ii) a test statistic calculated using the calculated allele ratios and the possible ratio of DNA to the expected distribution of the test statistic calculated using the predicted allele ratios and the possible ratio of DNA; ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment.

20. The method of any one of claims 3-19, wherein the sample comprises DNA from a mixture of one or more target cells and one or more non-target cells, and wherein the method further comprises:

creating a partition of possible ratios that range from a lower limit to an upper limit for the ratio of DNA from the one or more target cells to the total DNA in the sample;

enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment;

estimating, for each of the possible ratios of DNA in the partition and for each hypothesis, either (i) predicted allele ratios for the loci that are heterozygous in at least one cell given the possible ratio of DNA and the degree of overrepresentation specified by that hypothesis or (ii) an expected distribution of a test statistic calculated using the predicted allele ratios and the possible ratio of DNA;

calculating, for each of the possible ratios of DNA in the partition and for each hypothesis, the likelihood that the hypothesis is correct by comparing either (i) the calculated allele ratios to the predicted allele ratios, or (ii) a test statistic calculated using the calculated allele ratios and the possible ratio of DNA to the expected distribution of the test statistic calculated using the predicted allele ratios and the possible ratio of DNA; and

19. The method of any one of claims 3-18, wherein the sample comprises DNA from a mixture of one or more target cells and one or more non-target cells, and wherein the method further comprises:

enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment;

estimating, for each hypothesis, either (i) predicted allele ratios for the loci that are heterozygous in at least one cell given the degree of overrepresentation specified by that hypothesis or (ii) for one or more possible ratios of DNA, an expected distribution of a test statistic calculated using the predicted allele ratios and the possible ratio of DNA from the one or more target cells to the total DNA in the sample;

calculating, for each possible ratio of DNA, a data fit by comparing either (i) the calculated allele ratios to the predicted allele ratios, or (ii) a test statistic calculated using the calculated allele ratios and the possible ratio of DNA to the expected distribution of the test statistic calculated using the predicted allele ratios and the possible ratio of DNA; ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment.

20. The method of any one of claims 3-19, wherein the sample comprises DNA from a mixture of one or more target cells and one or more non-target cells, and wherein the method further comprises:

creating a partition of possible ratios that range from a lower limit to an upper limit for the ratio of DNA from the one or more target cells to the total DNA in the sample;

enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome segment;

estimating, for each of the possible ratios of DNA in the partition and for each hypothesis, either (i) predicted allele ratios for the loci that are heterozygous in at least one cell given the possible ratio of DNA and the degree of overrepresentation specified by that hypothesis or (ii) an expected distribution of a test statistic calculated using the predicted allele ratios and the possible ratio of DNA;

calculating, for each of the possible ratios of DNA in the partition and for each hypothesis, the likelihood that the hypothesis is correct by comparing either (i) the calculated allele ratios to the predicted allele ratios, or (ii) a test statistic calculated using the calculated allele ratios and the possible ratio of DNA to the expected distribution of the test statistic calculated using the predicted allele ratios and the possible ratio of DNA; and

determining the combined probability for each hypothesis by combining the probabilities of that hypothesis for each of the possible ratios in the partition; and

selecting the hypothesis with the greatest combined probability, thereby determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment.

21. The method of any one of claims 1-20, comprising determining that the first homologous chromosome segments are present in equal proportions if there is not an overrepresentation of the number of copies of the first homologous chromosome segment, and there is not an overrepresentation of the second homologous chromosome segment.

22. The method of any one of claims 3-21, wherein the expected allele ratio is 0.5 for biallelic loci.

23. The method of any one of claims 1-22, comprising applying a technique selected from the group consisting of maximum likelihood estimation, maximum a-posteriori estimation, Bayesian estimation, dynamic estimation, and expectation-maximization estimation to the obtained genetic data and the expected genetic data.

24. The method of any one of claims 1-23, wherein the sample comprises DNA from (i) one or more target cells, (ii) one or more non-target cells, or (iii) a mixture of one or more target cells and one or more non-target cells.

25. The method of claim 24, wherein the one or more target cells are cancer cell(s), and the one or more non-target cells are non-cancerous cell(s); and wherein the method comprises determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more of the cancer cells.

26. The method of claim 24, wherein the one or more target cells are genetically identical cancer cell(s), and the one or more non-target cells are non-cancerous cell(s); and wherein the method comprises determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of the cancer cell(s).

27. The method of claim 24, wherein the one or more target cells are genetically nonidentical

determining the combined probability for each hypothesis by combining the probabilities of that hypothesis for each of the possible ratios in the partition; and

selecting the hypothesis with the greatest combined probability, thereby determining the degree of overrepresentation of the number of copies of the first homologous chromosome segment.

21. The method of any one of claims 1-20, comprising determining that the first homologous chromosome segments are present in equal proportions if there is not an overrepresentation of the number of copies of the first homologous chromosome segment, and there is not an overrepresentation of the second homologous chromosome segment.

22. The method of any one of claims 3-21, wherein the expected allele ratio is 0.5 for biallelic loci.

23. The method of any one of claims 1-22, comprising applying a technique selected from the group consisting of maximum likelihood estimation, maximum a-posteriori estimation, Bayesian estimation, dynamic estimation, and expectation-maximization estimation to the obtained genetic data and the expected genetic data.

24. The method of any one of claims 1-23, wherein the sample comprises DNA from (i) one or more target cells, (ii) one or more non-target cells, or (iii) a mixture of one or more target cells and one or more non-target cells.

25. The method of claim 24, wherein the one or more target cells are cancer cell(s), and the one or more non-target cells are non-cancerous cell(s); and wherein the method comprises determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more of the cancer cells.

26. The method of claim 24, wherein the one or more target cells are genetically identical cancer cell(s), and the one or more non-target cells are non-cancerous cell(s); and wherein the method comprises determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of the cancer cell(s).

27. The method of claim 24, wherein the one or more target cells are genetically nonidentical

cancer cell(s), and the one or more non-target cells are non-cancerous cell(s); and wherein the method comprises determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more of the genetically non-identical cancer cells.

28. The method of claim 24, wherein the sample comprises cell-free DNA from a mixture of one or more cancer cells and one or more non-cancerous cells, and wherein the method comprises determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more of the cancer cells.

29. The method of any one of claims 1-28, wherein the measured genetic allelic data is obtained by (i) sequencing the DNA in the sample or (ii) amplifying DNA in the sample and then sequencing the amplified DNA.

30. The method of any one of claims 1-29, wherein the loci comprise single nucleotide polymorphisms (SNPs).

31. The method of claim of any one of claims 1-31, comprising determining whether one or more of the cells comprise a chromosomal condition selected from the group consisting of nullsomy, monosomy, disomy, trisomy, tetrasomy, deletion, duplication, mosaicism, and combinations thereof.

32. A report comprising a result from the method of any one of claims 1-31.

33. A method for determining a number of copies of a chromosome or chromosome segment in the genome of one or more cells from an individual, comprising:

obtaining genetic data at a set of polymorphic loci on the chromosome or chromosome segment in a sample of DNA from one or more cells from the individual by measuring the quantity of each allele at each locus;

calculating allele ratios for the loci that are heterozygous in at least one cell; wherein the allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus;

calculating a test statistic based on the magnitude of the difference between the calculated allele ratio and the expected allele ratio for each of the loci;

enumerating a set of one or more hypotheses specifying the number of copies of the chromosome

cancer cell(s), and the one or more non-target cells are non-cancerous cell(s); and wherein the method comprises determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more of the genetically non-identical cancer cells.

28. The method of claim 24, wherein the sample comprises cell-free DNA from a mixture of one or more cancer cells and one or more non-cancerous cells, and wherein the method comprises determining if there is an overrepresentation of the number of copies of the first homologous chromosome segment in the genome of one or more of the cancer cells.

29. The method of any one of claims 1-28, wherein the measured genetic allelic data is obtained by (i) sequencing the DNA in the sample or (ii) amplifying DNA in the sample and then sequencing the amplified DNA.

30. The method of any one of claims 1-29, wherein the loci comprise single nucleotide polymorphisms (SNPs).

31. The method of claim of any one of claims 1-31, comprising determining whether one or more of the cells comprise a chromosomal condition selected from the group consisting of nullsomy, monosomy, disomy, trisomy, tetrasomy, deletion, duplication, mosaicism, and combinations thereof.

32. A report comprising a result from the method of any one of claims 1-31.

33. A method for determining a number of copies of a chromosome or chromosome segment in the genome of one or more cells from an individual, comprising:

obtaining genetic data at a set of polymorphic loci on the chromosome or chromosome segment in a sample of DNA from one or more cells from the individual by measuring the quantity of each allele at each locus;

calculating allele ratios for the loci that are heterozygous in at least one cell; wherein the allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus;

calculating a test statistic based on the magnitude of the difference between the calculated allele ratio and the expected allele ratio for each of the loci;

enumerating a set of one or more hypotheses specifying the number of copies of the chromosome

or chromosome segment in the genome of one or more of the cells;

selecting which hypothesis is most likely based on the test statistic, thereby determining the number of copies of the chromosome or chromosome segment in the genome of one or more of the cells.

34. A method for determining a number of copies of a chromosome or chromosome segment in the genome of one or more from an individual, comprising:

obtaining genetic data at a set of polymorphic loci on the chromosome or chromosome segment in a mixed sample of cell-free DNA from the individual; wherein the mixed sample comprises cell-free DNA from two or more genetically different cells;

calculating allele ratios for the loci that are heterozygous in at least one cell; wherein the allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus;

calculating a test statistic based on the magnitude of the difference between the calculated allele ratio and the expected allele ratio for each of the loci;

enumerating a set of one or more hypotheses specifying the number of copies of the chromosome or chromosome segment in the genome of one or more of the cells;

selecting which hypothesis is most likely based on the test statistic, thereby determining the number of copies of the chromosome or chromosome segment in the genome of one or more of the cells.

35. The method of claim 33 or 34, comprising

selecting one of the hypotheses if the probability that the test statistic belongs to a distribution of the test statistic for that hypothesis is above an upper threshold;

rejecting the hypothesis if the probability that the test statistic belongs to the distribution of the test statistic for that hypothesis is below an lower threshold; or

neither selecting nor rejecting the hypothesis if the probability that the test statistic belongs to the distribution of the test statistic for that hypothesis is between the lower threshold and the upper threshold, or if the probability is not determined with sufficiently high confidence.

36. The method of any one of claims 33-35, wherein the test statistic is assumed to have a normal

or chromosome segment in the genome of one or more of the cells;

selecting which hypothesis is most likely based on the test statistic, thereby determining the number of copies of the chromosome or chromosome segment in the genome of one or more of the cells.

34. A method for determining a number of copies of a chromosome or chromosome segment in the genome of one or more from an individual, comprising:

obtaining genetic data at a set of polymorphic loci on the chromosome or chromosome segment in a mixed sample of cell-free DNA from the individual; wherein the mixed sample comprises cell-free DNA from two or more genetically different cells;

calculating allele ratios for the loci that are heterozygous in at least one cell; wherein the allele ratio for a particular locus is the measured quantity of one of the alleles divided by the total measured quantity of all the alleles for the locus;

calculating a test statistic based on the magnitude of the difference between the calculated allele ratio and the expected allele ratio for each of the loci;

enumerating a set of one or more hypotheses specifying the number of copies of the chromosome or chromosome segment in the genome of one or more of the cells;

selecting which hypothesis is most likely based on the test statistic, thereby determining the number of copies of the chromosome or chromosome segment in the genome of one or more of the cells.

35. The method of claim 33 or 34, comprising

selecting one of the hypotheses if the probability that the test statistic belongs to a distribution of the test statistic for that hypothesis is above an upper threshold;

rejecting the hypothesis if the probability that the test statistic belongs to the distribution of the test statistic for that hypothesis is below an lower threshold; or

neither selecting nor rejecting the hypothesis if the probability that the test statistic belongs to the distribution of the test statistic for that hypothesis is between the lower threshold and the upper threshold, or if the probability is not determined with sufficiently high confidence.

36. The method of any one of claims 33-35, wherein the test statistic is assumed to have a normal

distribution.

37. The method of any one of claims 33-36, wherein the test statistic A is calculated using the following formula

$A =$

wherein  $\delta_i$  is the magnitude of the difference between the calculated allele ratio and the expected allele ratio for the  $i$ th loci;

wherein  $\bar{a}$  is the mean value of  $a_i$ , and  
wherein  $s_a$  is the standard deviation of  $a_i$ .

38. The method of any one of claims 33-37, wherein the sample comprises DNA from (i) one or more target cells, (ii) one or more non-target cells, or (iii) a mixture of one or more target cells and one or more non-target cells.

39. The method of claim 38, wherein the sample comprises DNA from a mixture of one or more target cells and one or more non-target cells, and wherein the one or more target cells are cancer cell(s), and the one or more non-target cells are non-cancerous cell(s); and wherein the method comprises determining the number of copies of the chromosome or chromosome segment in the genome of one or more of the cancer cells.

40. The method of any one of claims 33-39, wherein the loci comprise SNPs.

41. The method of any one of claims 33-40, comprising determining whether one or more of the cells comprise a chromosomal condition selected from the group consisting of nullsomy, monosomy, disomy, trisomy, tetrasomy, deletion, duplication, mosaicism, and combinations thereof.

42. A report comprising a result from the method of any one of claims 33-41.

43. A method for determining a number of copies of a chromosome segment of interest in the genome of a fetus, comprising:

obtaining phased genetic data for at least one biological parent of the fetus, wherein the phased

distribution.

37. The method of any one of claims 33-36, wherein the test statistic A is calculated using the following formula

$A =$

wherein  $\delta_i$  is the magnitude of the difference between the calculated allele ratio and the expected allele ratio for the  $i$ th loci;

wherein  $\bar{a}$  is the mean value of  $a_i$ , and  
wherein  $s_a$  is the standard deviation of  $a_i$ .

38. The method of any one of claims 33-37, wherein the sample comprises DNA from (i) one or more target cells, (ii) one or more non-target cells, or (iii) a mixture of one or more target cells and one or more non-target cells.

39. The method of claim 38, wherein the sample comprises DNA from a mixture of one or more target cells and one or more non-target cells, and wherein the one or more target cells are cancer cell(s), and the one or more non-target cells are non-cancerous cell(s); and wherein the method comprises determining the number of copies of the chromosome or chromosome segment in the genome of one or more of the cancer cells.

40. The method of any one of claims 33-39, wherein the loci comprise SNPs.

41. The method of any one of claims 33-40, comprising determining whether one or more of the cells comprise a chromosomal condition selected from the group consisting of nullsomy, monosomy, disomy, trisomy, tetrasomy, deletion, duplication, mosaicism, and combinations thereof.

42. A report comprising a result from the method of any one of claims 33-41.

43. A method for determining a number of copies of a chromosome segment of interest in the genome of a fetus, comprising:

obtaining phased genetic data for at least one biological parent of the fetus, wherein the phased

genetic data comprises the identity of the allele present for each locus in a set of polymorphic loci on a first homologous chromosome segment and a second homologous chromosome segment in a pair of homologous chromosome segments that comprises the chromosome segment of interest;

obtaining genetic data at the set of polymorphic loci on the chromosome segment of interest in a mixed sample of DNA comprising fetal DNA and maternal DNA from the mother of the fetus by measuring the quantity of each allele at each locus;

enumerating a set of one or more hypotheses specifying, for one or both parents, the number of copies of the first homologous chromosome segment or portion thereof from the parent in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion thereof from the parent in the genome of the fetus, and the total number of copies of the chromosome segment of interest present in the genome of the fetus;

calculating, for each of the hypotheses, expected genetic data for the plurality of loci in the mixed sample from the obtained phased genetic data from the parent(s);

calculating, on a computer, the data fit between the obtained genetic data of the mixed sample and the expected genetic data for the mixed sample;

ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the number of copies of the chromosome segment of interest in the genome of the fetus.

44. A method for determining a number of copies of a chromosome or chromosome segment of interest in the genome of a fetus, comprising:

obtaining phased genetic data for at least one biological parent of the fetus, wherein the phased genetic data comprises the identity of the allele present for each locus in a set of polymorphic loci on a first homologous chromosome segment and a second homologous chromosome segment in the parent,

obtaining genetic data at the set of polymorphic loci on the chromosome or chromosome segment in a mixed sample of DNA comprising fetal DNA and maternal DNA from the mother of the fetus by measuring the quantity of each allele at each locus;

enumerating a set of one or more hypotheses specifying the number of copies of the chromosome or chromosome segment of interest present in the genome of the fetus;

creating, on a computer, for each of the hypotheses, a probability distribution of the expected quantity of each allele at each of the plurality of loci in mixed sample from the (i) the obtained phased

genetic data comprises the identity of the allele present for each locus in a set of polymorphic loci on a first homologous chromosome segment and a second homologous chromosome segment in a pair of homologous chromosome segments that comprises the chromosome segment of interest;

obtaining genetic data at the set of polymorphic loci on the chromosome segment of interest in a mixed sample of DNA comprising fetal DNA and maternal DNA from the mother of the fetus by measuring the quantity of each allele at each locus;

enumerating a set of one or more hypotheses specifying, for one or both parents, the number of copies of the first homologous chromosome segment or portion thereof from the parent in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion thereof from the parent in the genome of the fetus, and the total number of copies of the chromosome segment of interest present in the genome of the fetus;

calculating, for each of the hypotheses, expected genetic data for the plurality of loci in the mixed sample from the obtained phased genetic data from the parent(s);

calculating, on a computer, the data fit between the obtained genetic data of the mixed sample and the expected genetic data for the mixed sample;

ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the number of copies of the chromosome segment of interest in the genome of the fetus.

44. A method for determining a number of copies of a chromosome or chromosome segment of interest in the genome of a fetus, comprising:

obtaining phased genetic data for at least one biological parent of the fetus, wherein the phased genetic data comprises the identity of the allele present for each locus in a set of polymorphic loci on a first homologous chromosome segment and a second homologous chromosome segment in the parent,

obtaining genetic data at the set of polymorphic loci on the chromosome or chromosome segment in a mixed sample of DNA comprising fetal DNA and maternal DNA from the mother of the fetus by measuring the quantity of each allele at each locus;

enumerating a set of one or more hypotheses specifying the number of copies of the chromosome or chromosome segment of interest present in the genome of the fetus;

creating, on a computer, for each of the hypotheses, a probability distribution of the expected quantity of each allele at each of the plurality of loci in mixed sample from the (i) the obtained phased

genetic data from the parent(s) and (ii) the probability of one or more crossovers that may have occurred during the formation of a gamete that contributed a copy of the chromosome or chromosome segment of interest to the fetus;

calculating, on a computer, a fit, for each of the hypotheses, between (1) the obtained genetic data of the mixed sample and (2) the probability distribution of the expected quantity of each allele at each of the plurality of loci in mixed sample for that hypothesis; ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the number of copies of the chromosome segment of interest in the genome of the fetus.

45. The method of claim 43 or 44, comprising: obtaining

phased genetic data for the mother of the fetus;

enumerating a set of one or more hypotheses specifying the number of copies of the first homologous chromosome segment or portion thereof from the mother in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion thereof from the mother in the genome of the fetus, and the total number of copies of the chromosome segment of interest present in the genome of the fetus; and

calculating, for each of the hypotheses, expected genetic data for the plurality of loci in the mixed sample from the obtained phased genetic data from the mother.

46. The method of claim 45, wherein the expected genetic data for each of the hypotheses comprises the identity and an amount of one or more alleles at each locus in the plurality of loci from the maternal DNA and fetal DNA in the mixed sample; and wherein calculating expected genetic data comprises:

determining a fraction of fetal DNA and a fraction of maternal DNA in the mixed sample;

calculating, for each locus in the plurality of loci, the expected amount of one or more of the alleles for that locus in the maternal DNA in the mixed sample using the identity of the allele(s) present at that locus in the obtained phased genetic data of the mother and the fraction of maternal DNA in the mixed sample; and

calculating, for each locus in the plurality of loci for each hypothesis, the expected amount of one or more of the alleles for that locus in the fetal DNA inherited from the mother in the mixed sample

genetic data from the parent(s) and (ii) the probability of one or more crossovers that may have occurred during the formation of a gamete that contributed a copy of the chromosome or chromosome segment of interest to the fetus;

calculating, on a computer, a fit, for each of the hypotheses, between (1) the obtained genetic data of the mixed sample and (2) the probability distribution of the expected quantity of each allele at each of the plurality of loci in mixed sample for that hypothesis; ranking one or more of the hypotheses according to the data fit; and selecting the hypothesis that is ranked the highest, thereby determining the number of copies of the chromosome segment of interest in the genome of the fetus.

45. The method of claim 43 or 44, comprising: obtaining phased genetic data for the mother of the fetus;

enumerating a set of one or more hypotheses specifying the number of copies of the first homologous chromosome segment or portion thereof from the mother in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion thereof from the mother in the genome of the fetus, and the total number of copies of the chromosome segment of interest present in the genome of the fetus; and

calculating, for each of the hypotheses, expected genetic data for the plurality of loci in the mixed sample from the obtained phased genetic data from the mother.

46. The method of claim 45, wherein the expected genetic data for each of the hypotheses comprises the identity and an amount of one or more alleles at each locus in the plurality of loci from the maternal DNA and fetal DNA in the mixed sample; and wherein calculating expected genetic data comprises:

determining a fraction of fetal DNA and a fraction of maternal DNA in the mixed sample;

calculating, for each locus in the plurality of loci, the expected amount of one or more of the alleles for that locus in the maternal DNA in the mixed sample using the identity of the allele(s) present at that locus in the obtained phased genetic data of the mother and the fraction of maternal DNA in the mixed sample; and

calculating, for each locus in the plurality of loci for each hypothesis, the expected amount of one or more of the alleles for that locus in the fetal DNA inherited from the mother in the mixed sample

using the identity of the allele present at that locus in the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, the number of copies of the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, and the fraction of fetal DNA in the mixed sample.

47. The method of claim 45, wherein the expected genetic data for each of the hypotheses comprises the identity and an amount of one or more alleles at each locus in the plurality of loci from the maternal DNA and fetal DNA in the mixed sample; and wherein calculating expected genetic data comprises:

determining a fraction of fetal DNA and a fraction of maternal DNA in the mixed sample;

calculating, for each locus in the plurality of loci, the expected amount of one or more of the alleles for that locus in the maternal DNA in the mixed sample using the identity of the allele(s) present at that locus in the obtained phased genetic data of the mother and the fraction of maternal DNA in the mixed sample; and

calculating, for each locus in the plurality of loci for each hypothesis, the expected amount of one or more of the alleles for that locus in the fetal DNA inherited from the mother in the mixed sample using the identity of the allele present at that locus in the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, the number of copies of the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, the identity of one or more possible alleles at that locus in the first or second homologous chromosome segment from the father that is specified by the hypothesis to have been inherited by the fetus, the number of copies of the first or second homologous chromosome segment from the father that is specified by the hypothesis to have been inherited by the fetus, and the fraction of fetal DNA in the mixed sample.

48. The method of claim 43 or 44, comprising:

obtaining phased genetic data for both the mother and father of the fetus;

enumerating a set of one or more hypotheses specifying the number of copies of the first homologous chromosome segment or portion thereof from the mother in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion thereof from

using the identity of the allele present at that locus in the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, the number of copies of the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, and the fraction of fetal DNA in the mixed sample.

47. The method of claim 45, wherein the expected genetic data for each of the hypotheses comprises the identity and an amount of one or more alleles at each locus in the plurality of loci from the maternal DNA and fetal DNA in the mixed sample; and wherein calculating expected genetic data comprises:

determining a fraction of fetal DNA and a fraction of maternal DNA in the mixed sample;

calculating, for each locus in the plurality of loci, the expected amount of one or more of the alleles for that locus in the maternal DNA in the mixed sample using the identity of the allele(s) present at that locus in the obtained phased genetic data of the mother and the fraction of maternal DNA in the mixed sample; and

calculating, for each locus in the plurality of loci for each hypothesis, the expected amount of one or more of the alleles for that locus in the fetal DNA inherited from the mother in the mixed sample using the identity of the allele present at that locus in the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, the number of copies of the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, the identity of one or more possible alleles at that locus in the first or second homologous chromosome segment from the father that is specified by the hypothesis to have been inherited by the fetus, the number of copies of the first or second homologous chromosome segment from the father that is specified by the hypothesis to have been inherited by the fetus, and the fraction of fetal DNA in the mixed sample.

48. The method of claim 43 or 44, comprising:

obtaining phased genetic data for both the mother and father of the fetus;

enumerating a set of one or more hypotheses specifying the number of copies of the first homologous chromosome segment or portion thereof from the mother in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion thereof from

the mother in the genome of the fetus, the number of copies of the first homologous chromosome segment or portion thereof from the father in the genome of the fetus, the number of copies of the second homologous chromosome segment or portion thereof from the father in the genome of the fetus, and the total number of copies of the chromosome segment of interest present in the genome of the fetus; and

calculating, for each of the hypotheses, expected genetic data for the plurality of loci in the mixed sample from the obtained phased genetic data from the mother and obtained phased genetic data from the father.

49. The method of claim 48, wherein the expected genetic data for each of the hypotheses comprises the identity and an amount of one or more alleles at each locus in the plurality of loci from the maternal DNA and fetal DNA in the mixed sample; and wherein calculating expected genetic data comprises:

determining a fraction of fetal DNA and a fraction of maternal DNA in the mixed sample;

calculating, for each locus in the plurality of loci, the expected amount of one or more of the alleles for that locus in the maternal DNA in the mixed sample using the identity of the allele(s) present at that locus in the obtained phased genetic data of the mother and the fraction of maternal DNA in the mixed sample; and

calculating, for each locus in the plurality of loci for each hypothesis, the expected amount of one or more of the alleles for that locus in the fetal DNA in the mixed sample using the identity of the allele present at that locus in the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, the number of copies of the first or second homologous chromosome segment from the mother that is specified by the hypothesis to have been inherited by the fetus, the identity of the allele present at that locus in the first or second homologous chromosome segment from the father that is specified by the hypothesis to have been inherited by the fetus, the number of copies of the first or second homologous chromosome segment from the father that is specified by the hypothesis to have been inherited by the fetus, and the fraction of fetal DNA in the mixed sample.

50. The method of any one of claims 43-49, comprising:

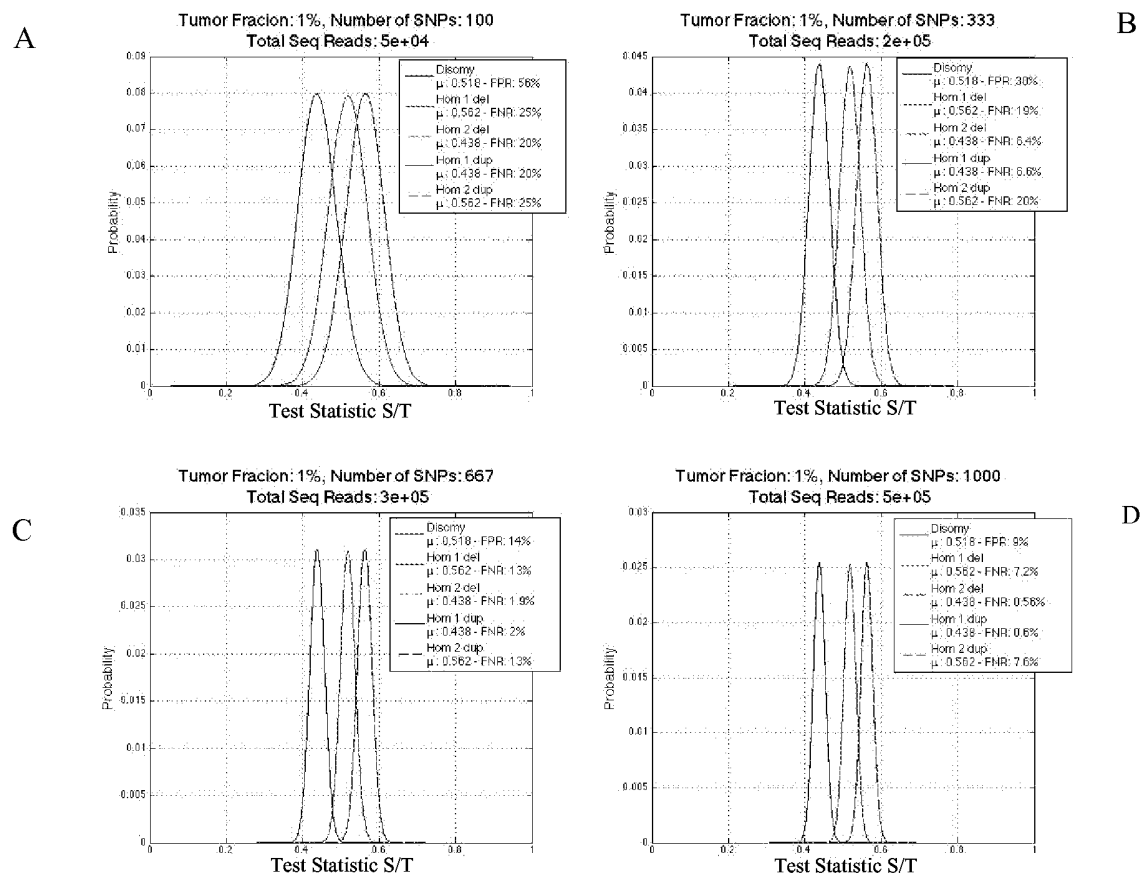
calculating, for each of the hypotheses, a probability distribution of expected genetic data for the plurality of loci in the mixed sample from the obtained phased genetic data from the parent(s); and

## **ABSTRACT**

**[00454]** The invention provides methods of detecting deletions and duplications of chromosome segments or entire chromosomes. In some aspects, the invention features methods for diagnosing, prognosticating, preventing, delaying, stabilizing, or treating a disease or disorder in a mammal, such as cancer or a chromosomal abnormality in a gestating fetus.

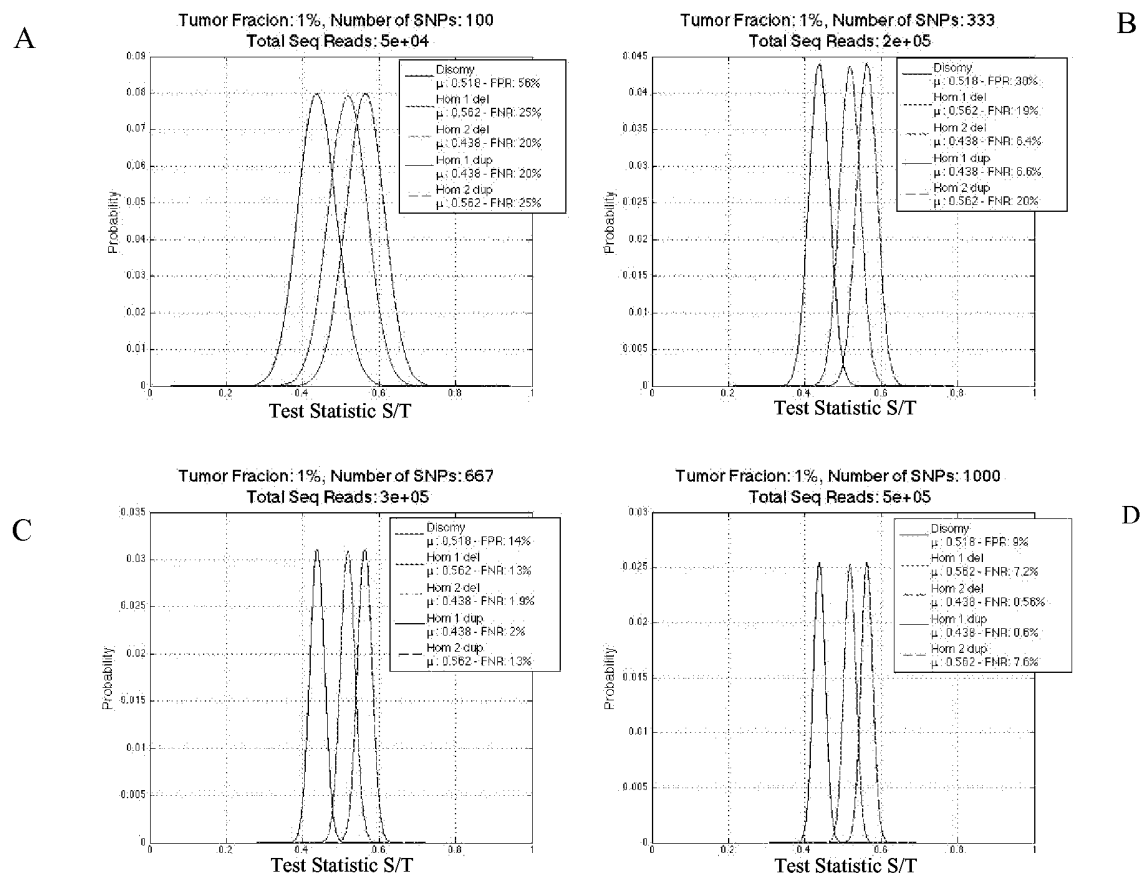
Figures 1A-1D

DOR of 500 and Tumor Fraction of 1%:



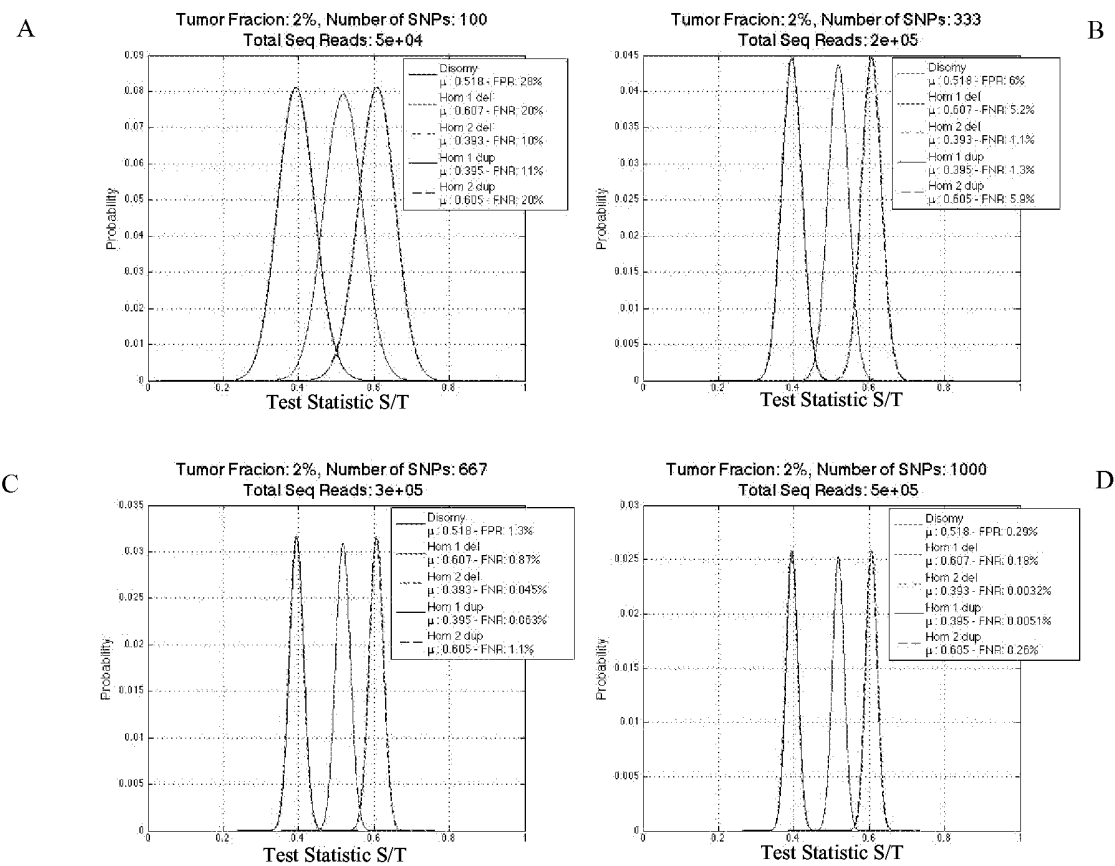
Figures 1A-1D

DOR of 500 and Tumor Fraction of 1%:



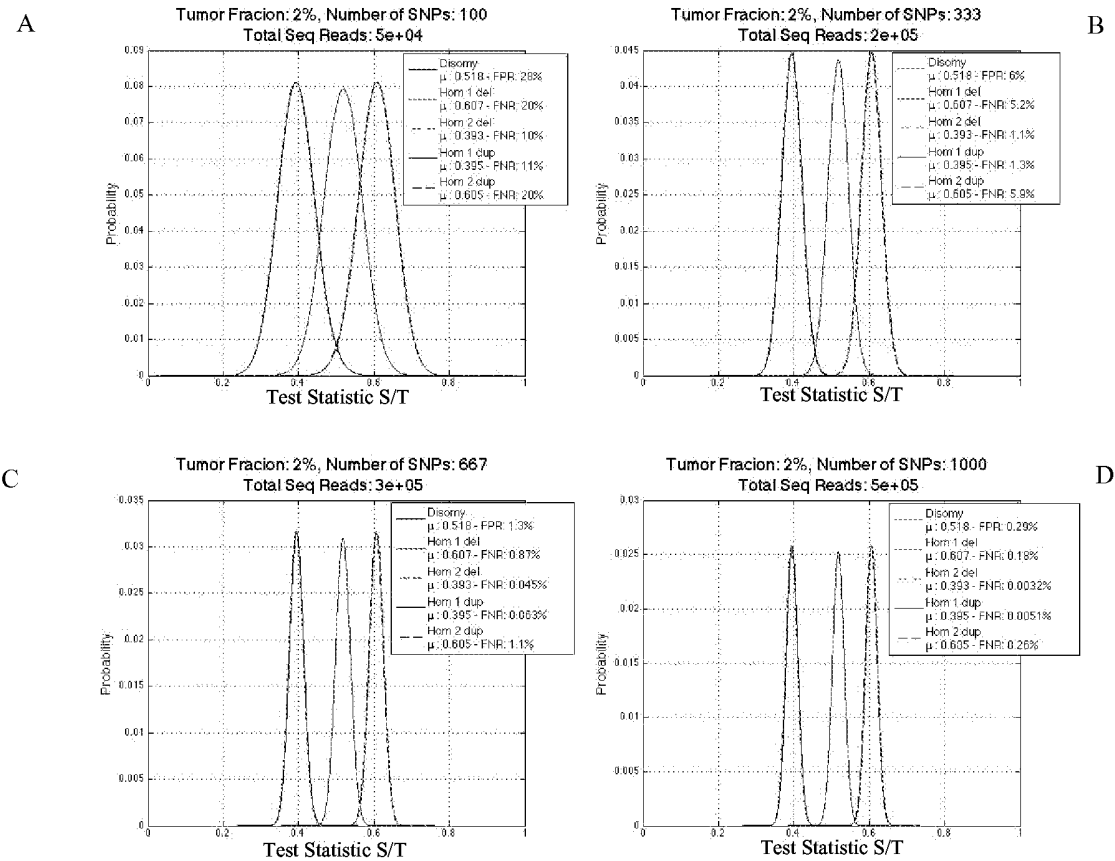
Figures 2A-2D

DOR of 500 and Tumor Fraction of 2%:



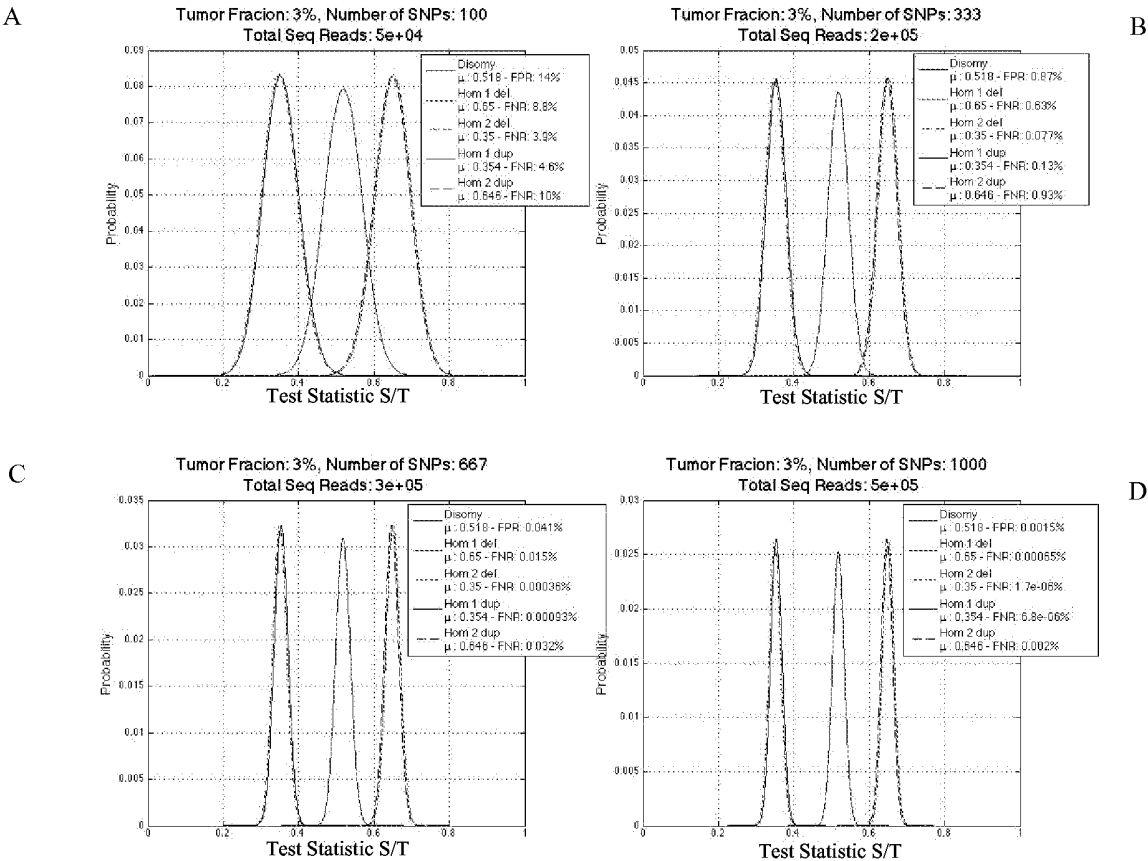
Figures 2A-2D

DOR of 500 and Tumor Fraction of 2%:



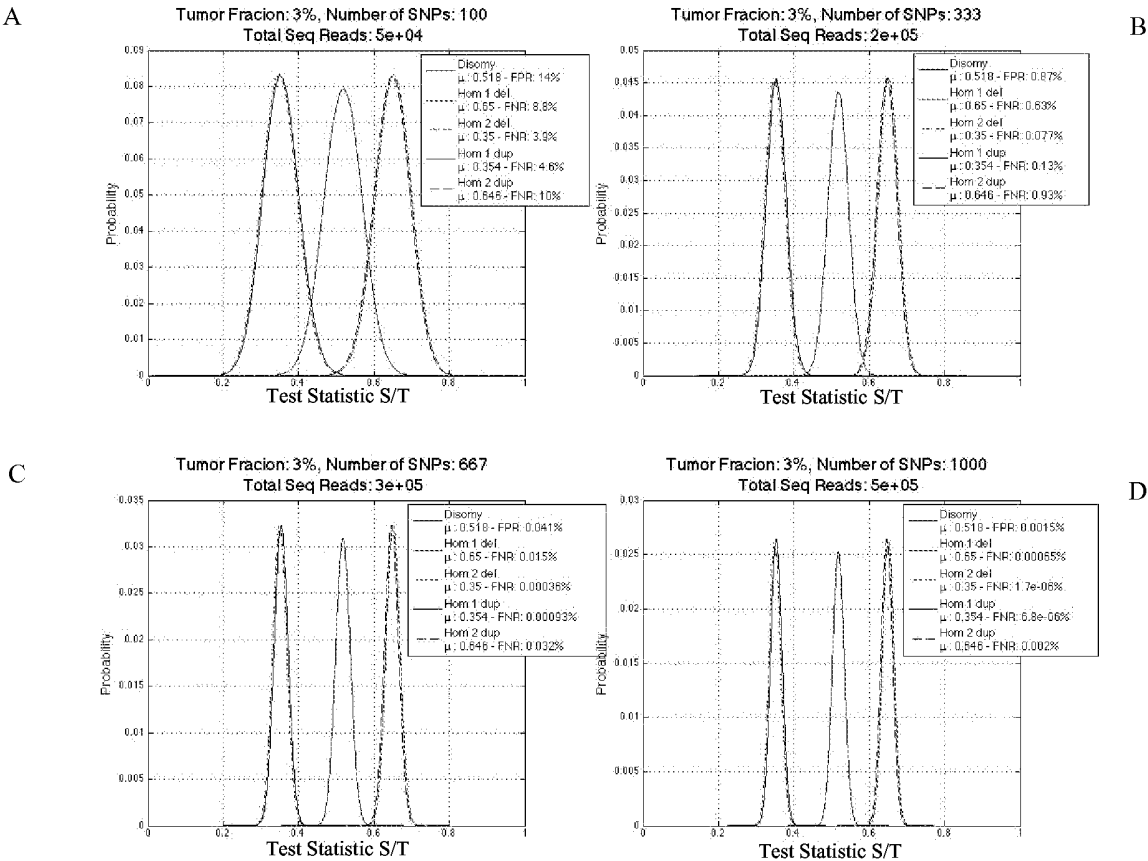
Figures 3A-3D

DOR of 500 and Tumor Fraction of 3%:



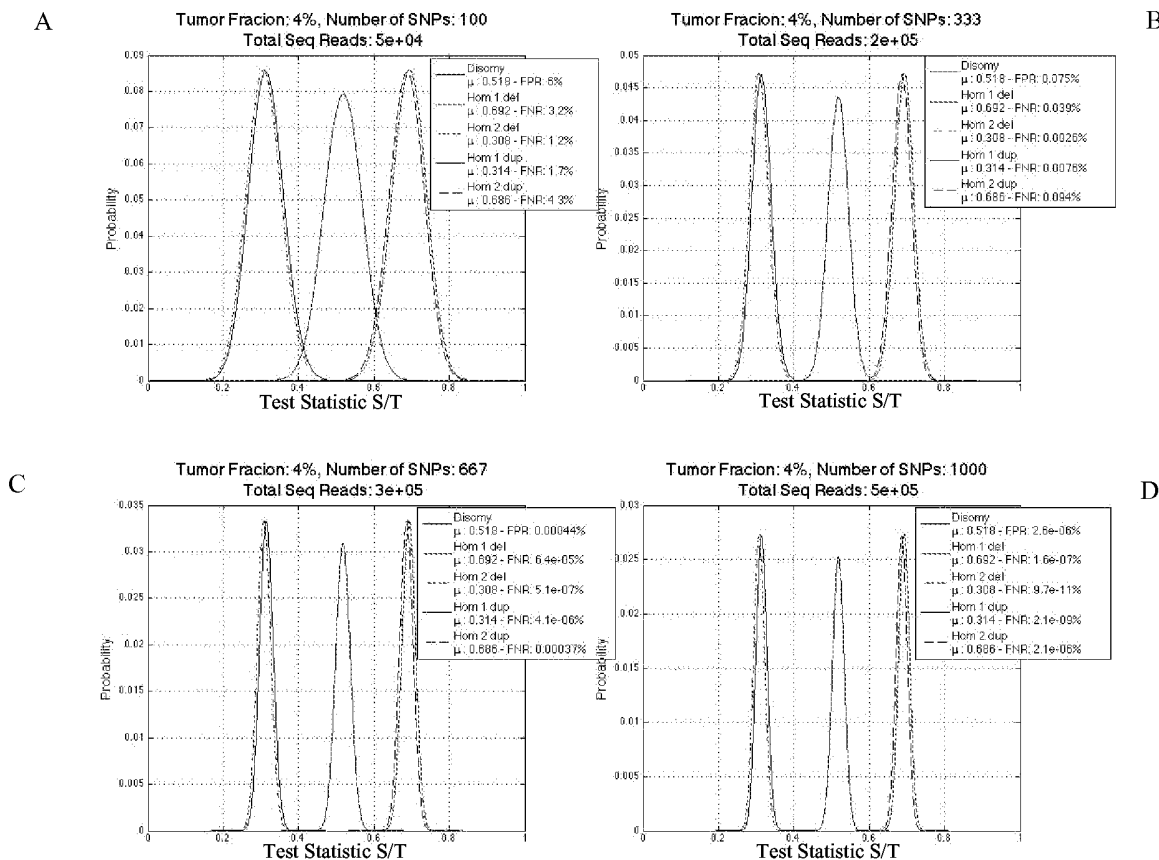
Figures 3A-3D

DOR of 500 and Tumor Fraction of 3%:



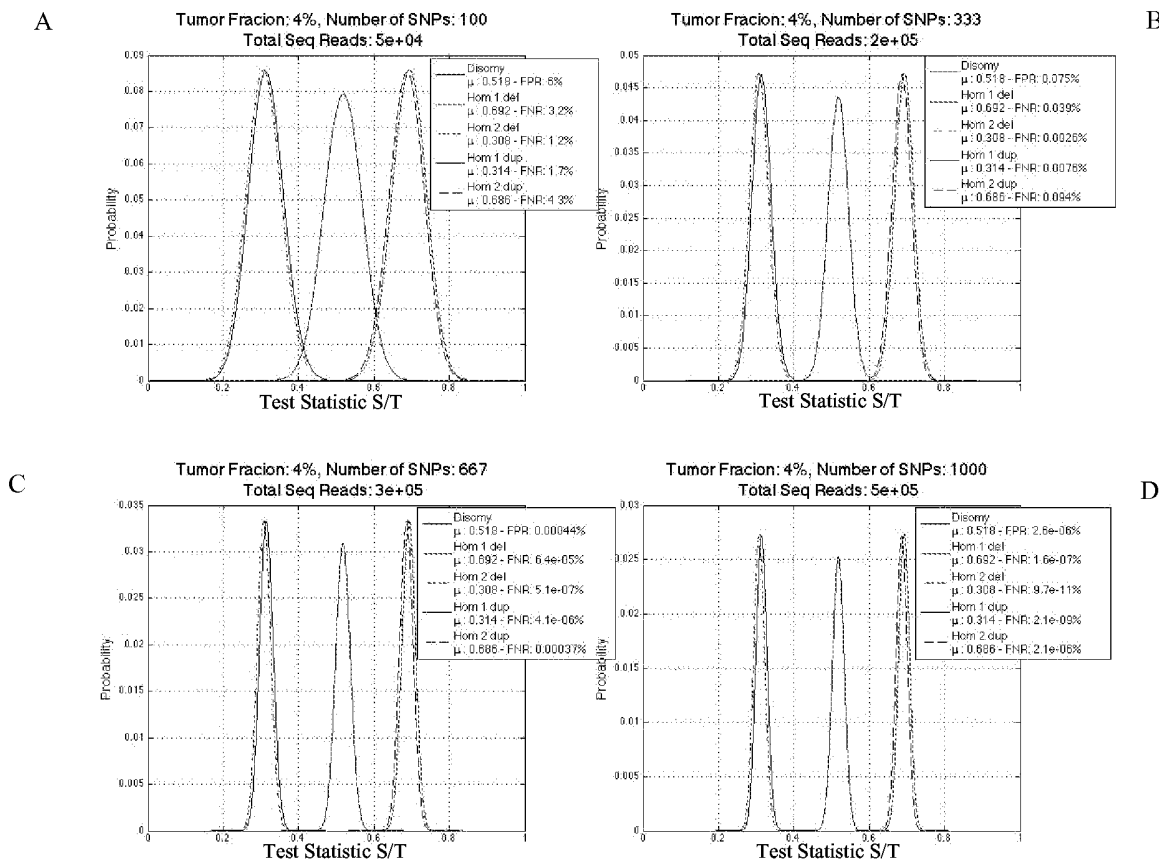
Figures 4A-4D

DOR of 500 and Tumor Fraction of 4%:



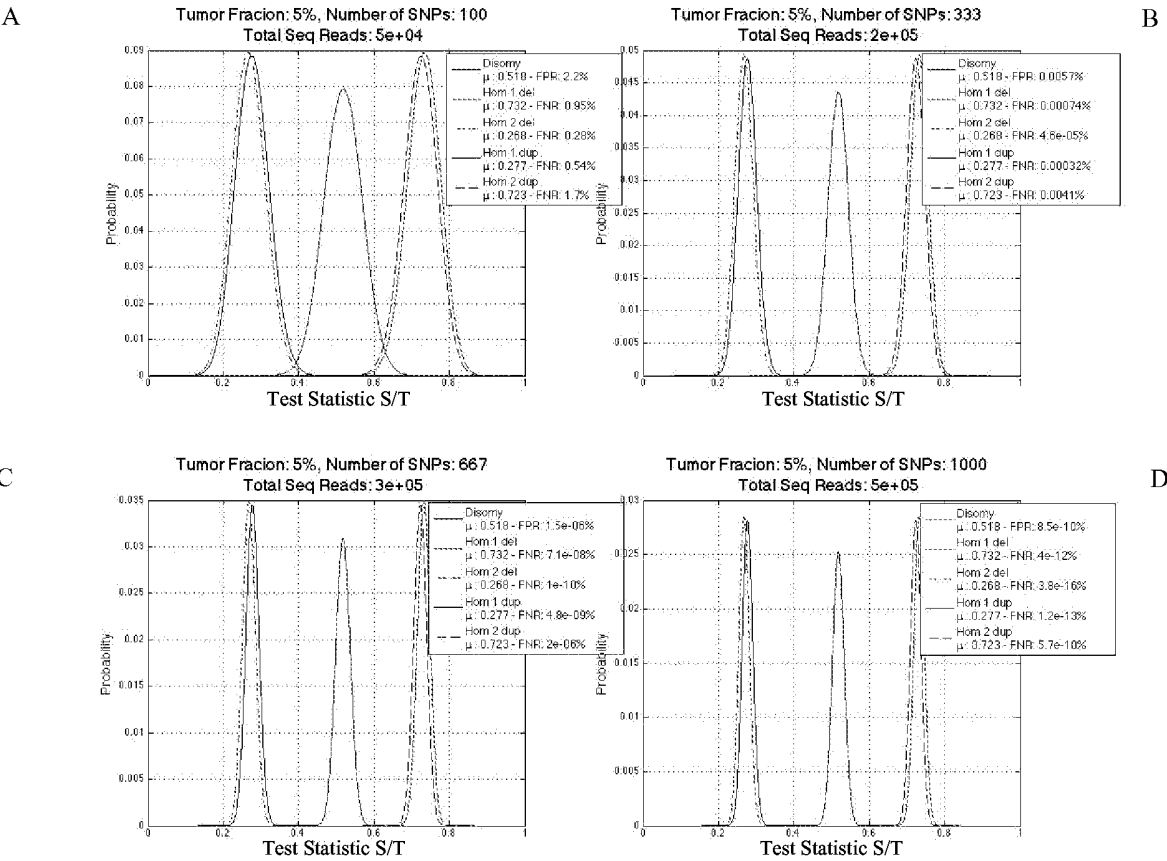
Figures 4A-4D

DOR of 500 and Tumor Fraction of 4%:



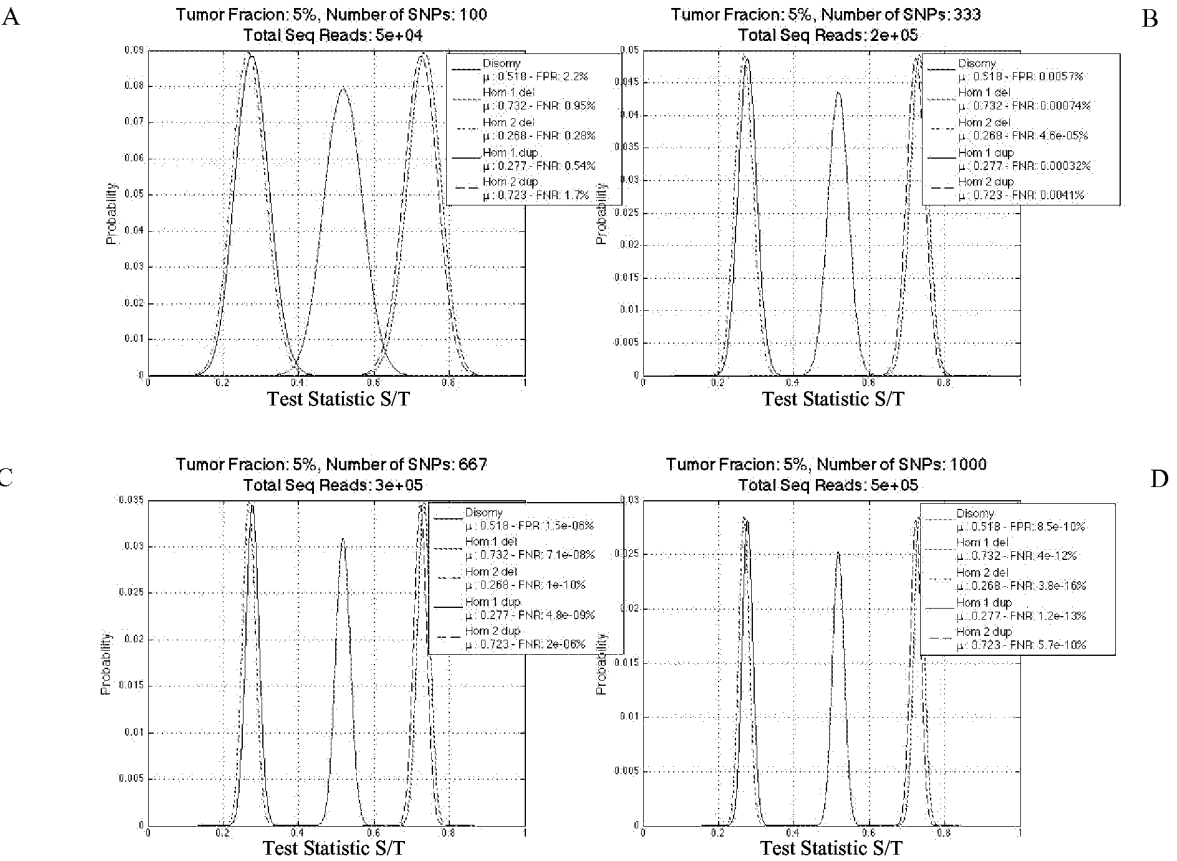
Figures 5A-5D

DOR of 500 and Tumor Fraction of 5%:



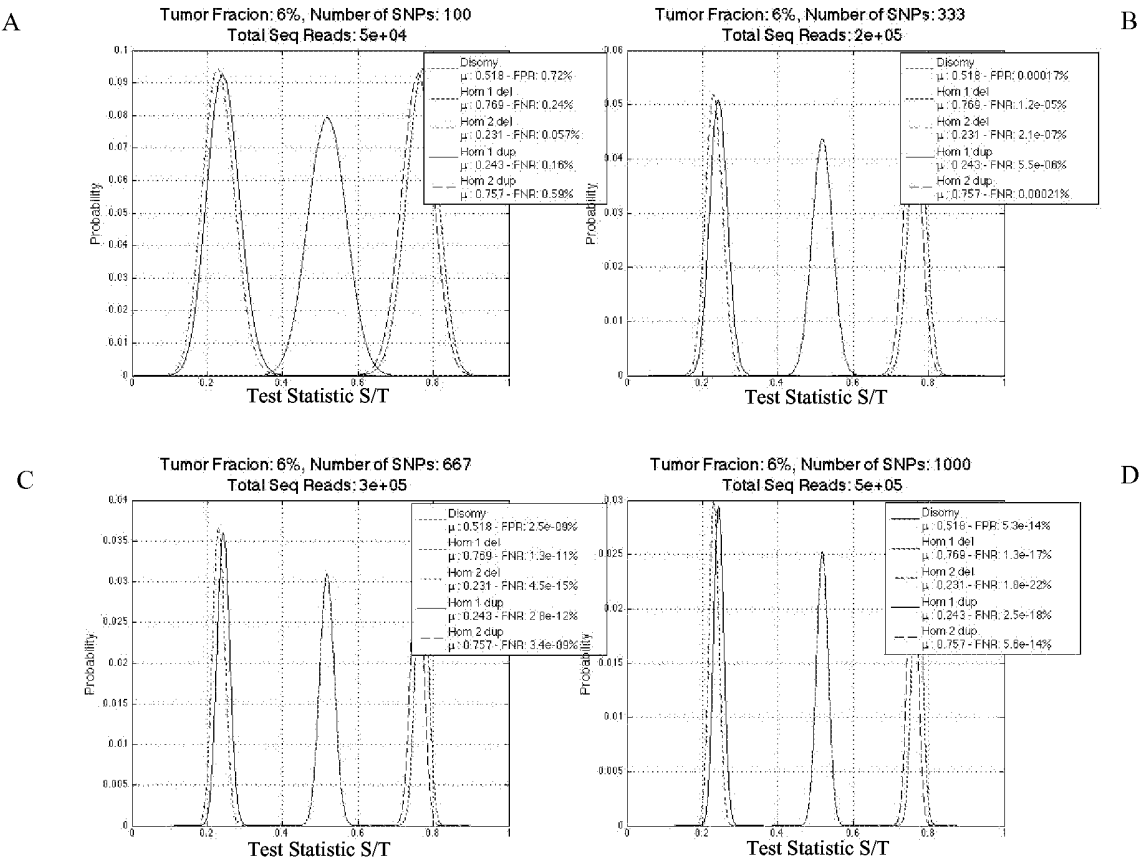
Figures 5A-5D

DOR of 500 and Tumor Fraction of 5%:



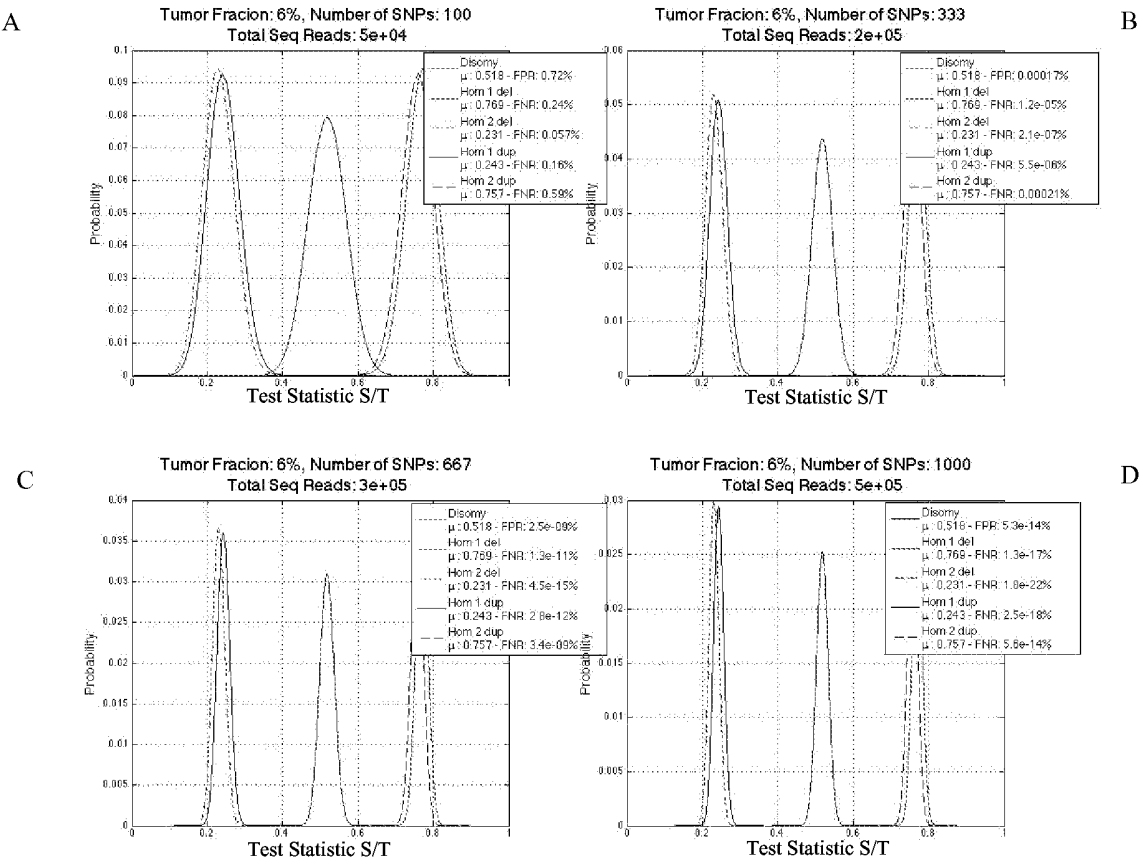
Figures 6A-6D

DOR of 500 and Tumor Fraction of 6%:



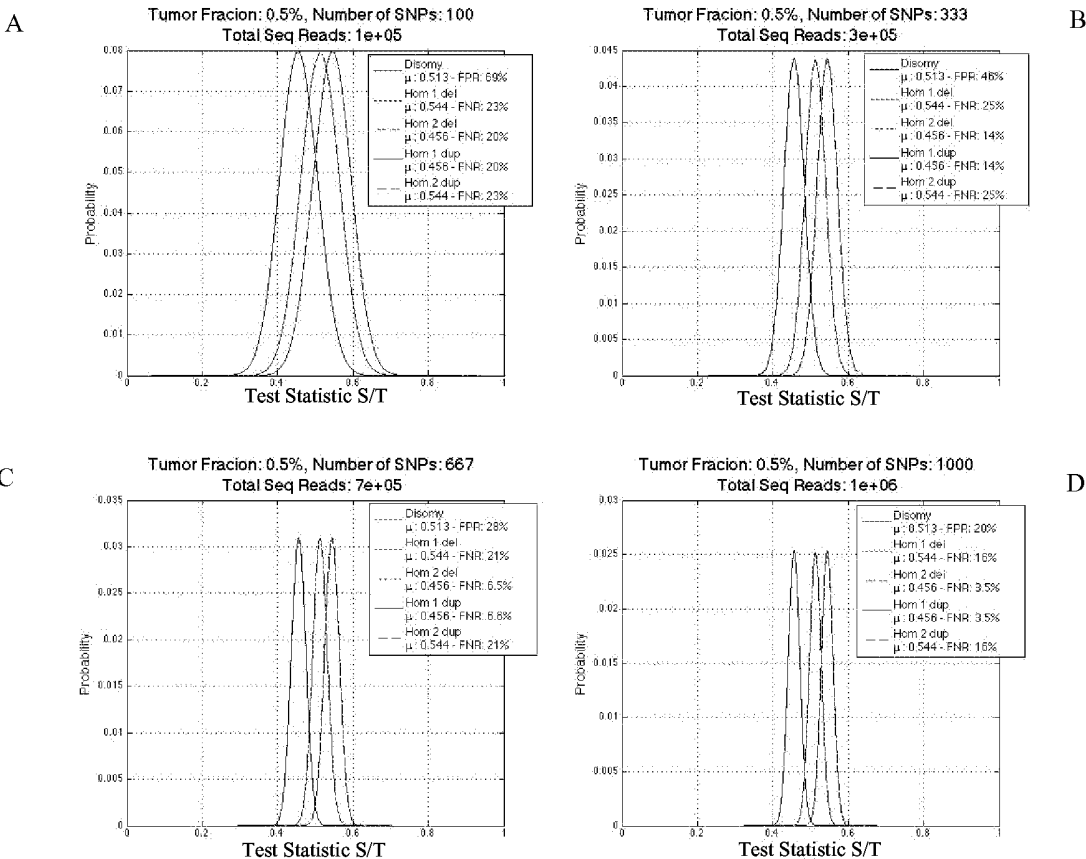
Figures 6A-6D

DOR of 500 and Tumor Fraction of 6%:



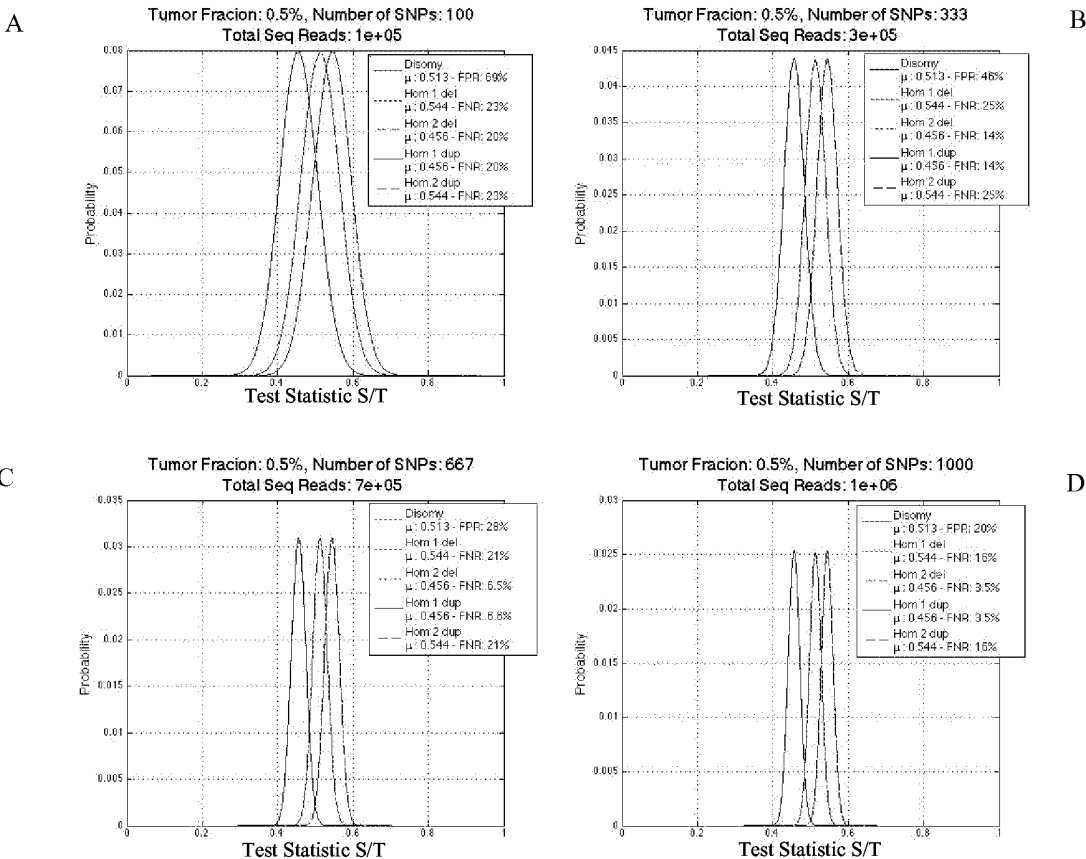
Figures 7A-7D

DOR of 1000 and Tumor Fraction of 0.5%:



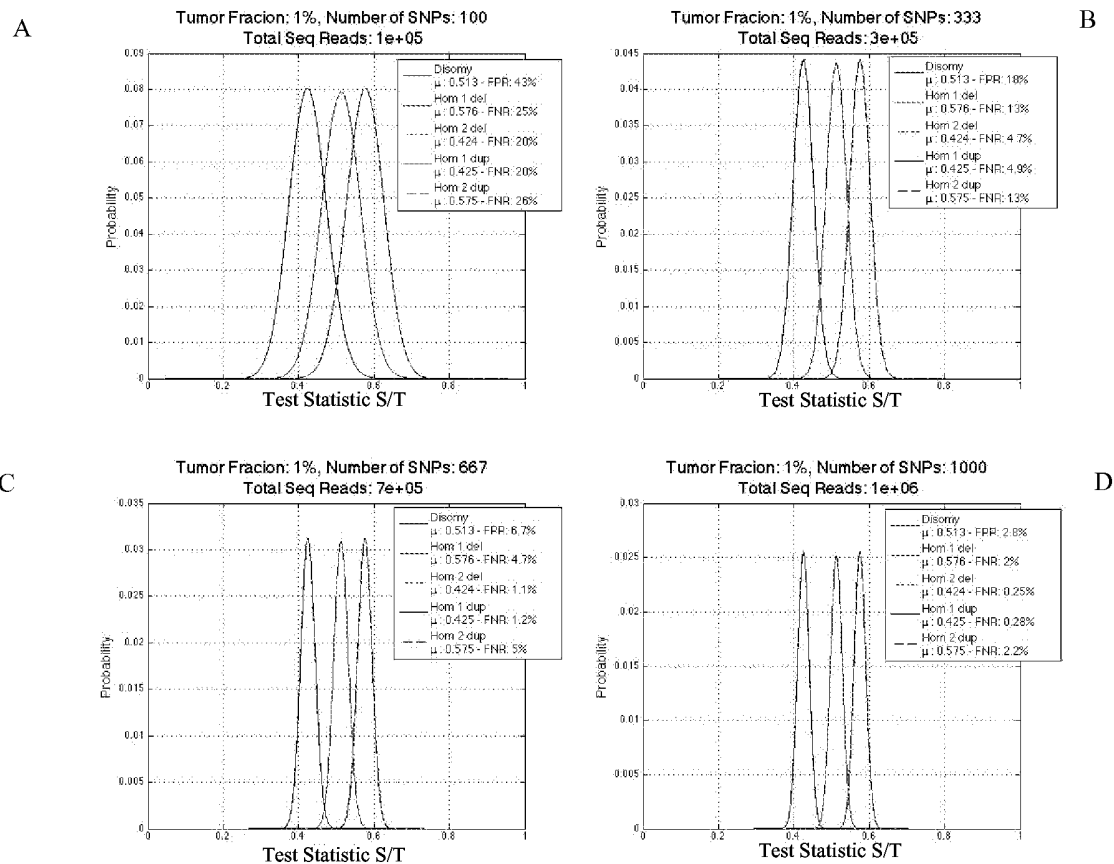
Figures 7A-7D

DOR of 1000 and Tumor Fraction of 0.5%:



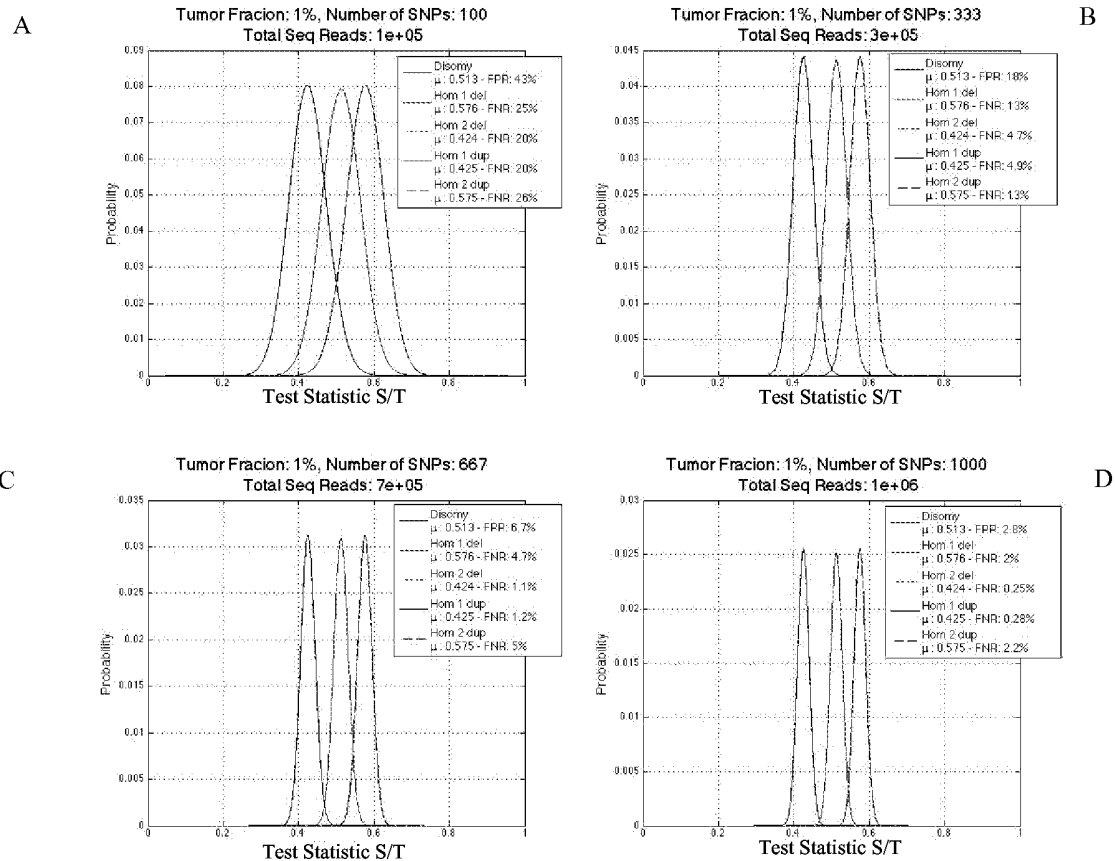
Figures 8A-8D

DOR of 1000 and Tumor Fraction of 1%:



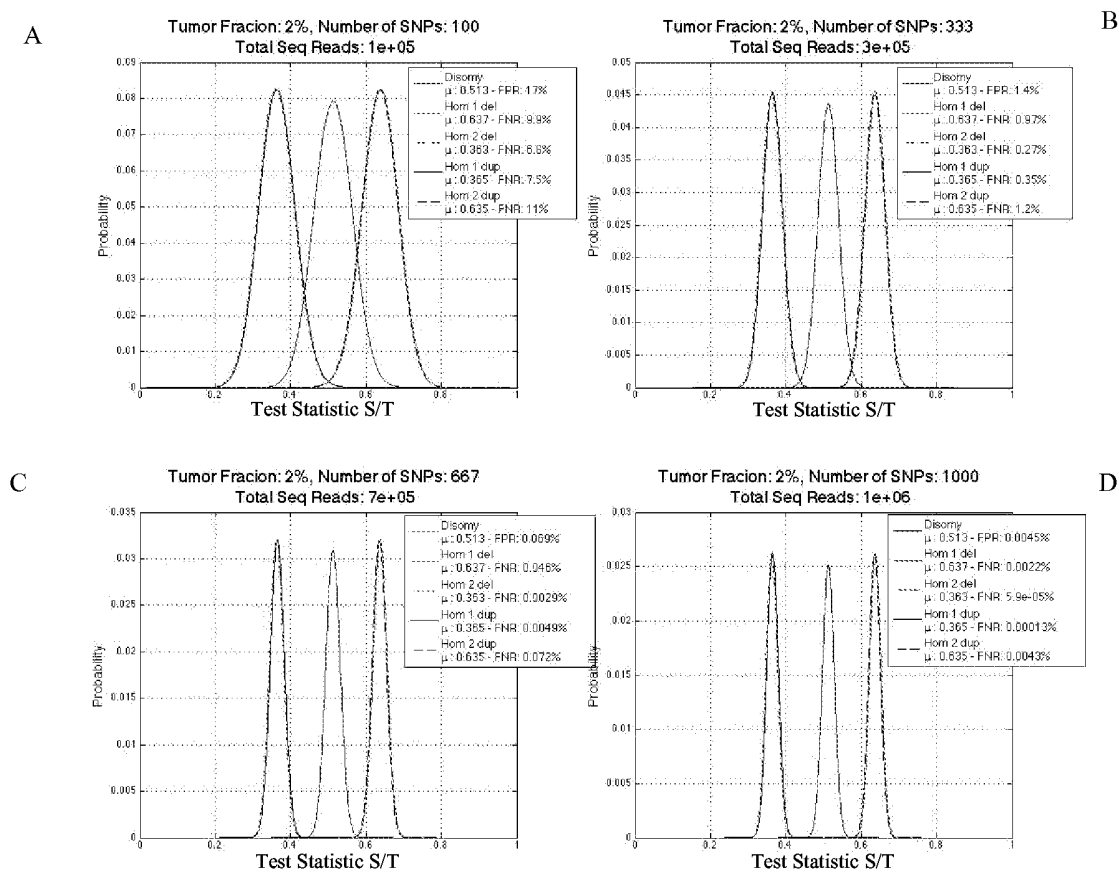
Figures 8A-8D

DOR of 1000 and Tumor Fraction of 1%:



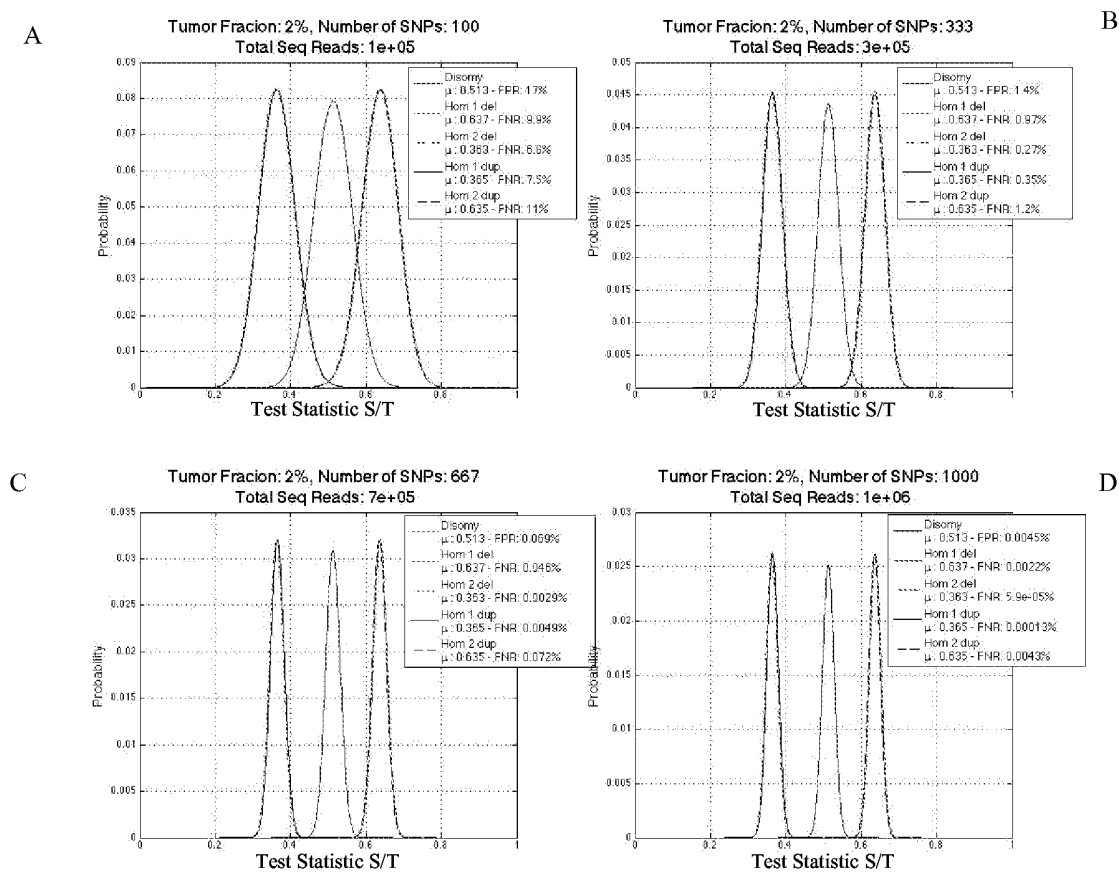
Figures 9A-9D

DOR of 1000 and Tumor Fraction of 2%:



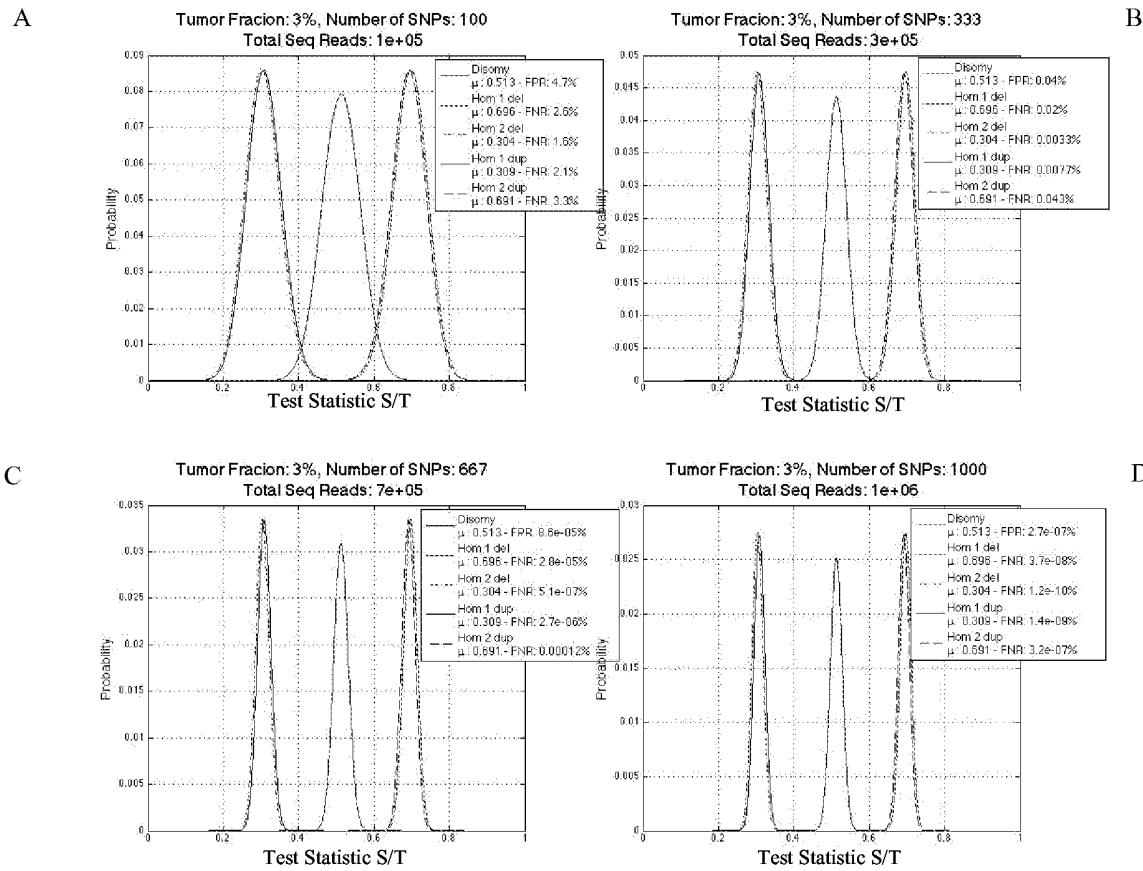
Figures 9A-9D

DOR of 1000 and Tumor Fraction of 2%:



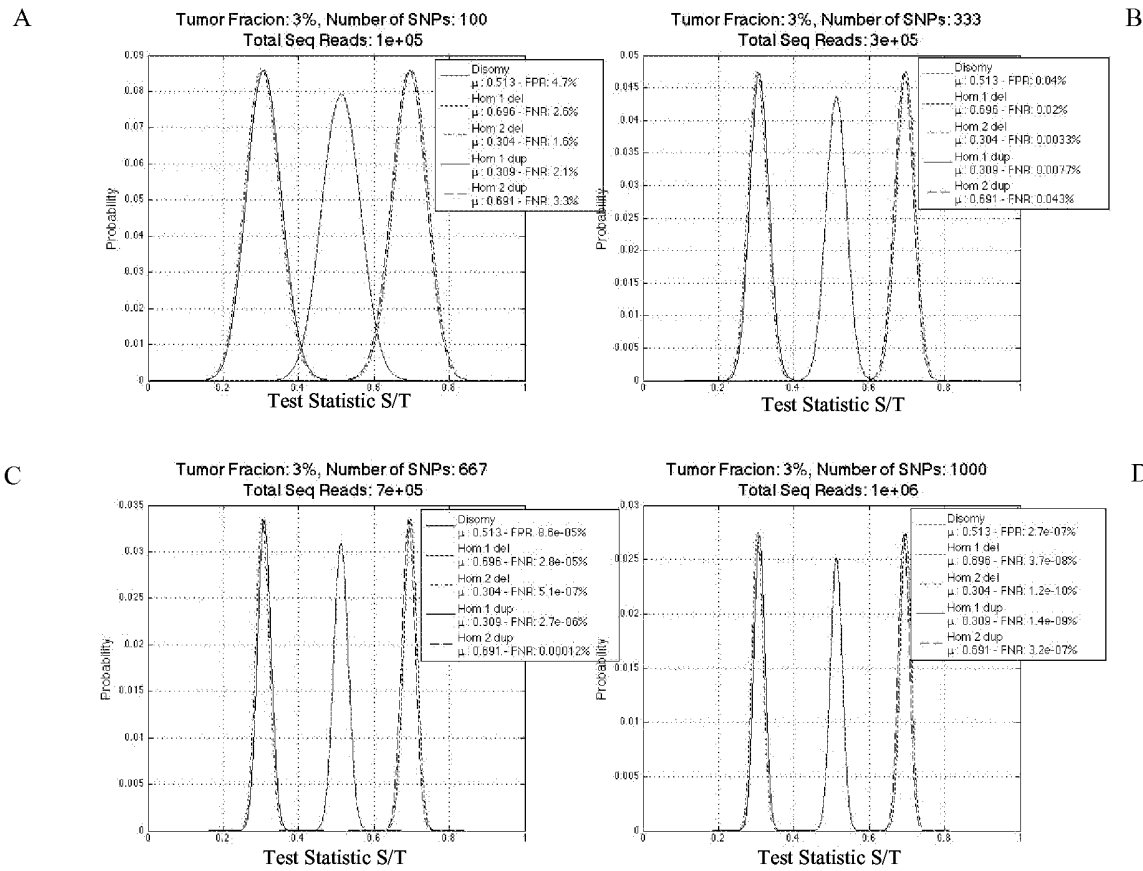
Figures 10A-10D

DOR of 1000 and Tumor Fraction of 3%:



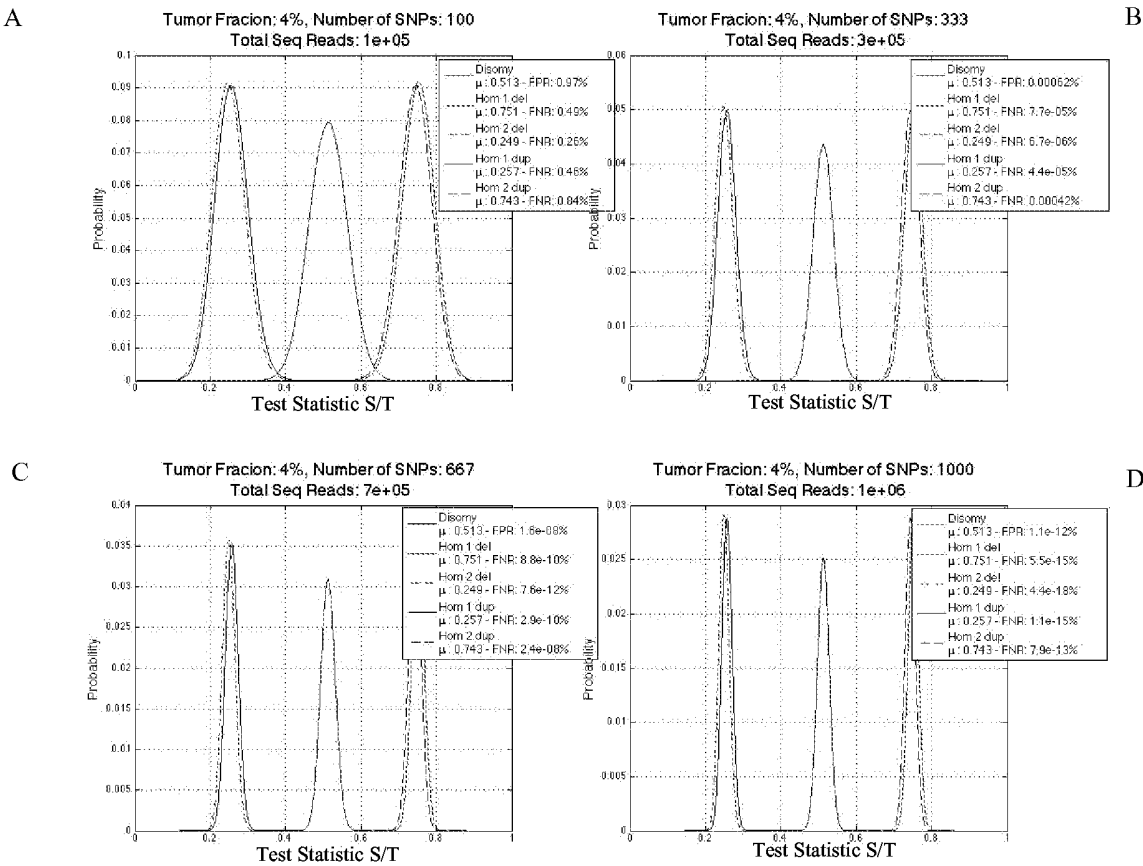
Figures 10A-10D

DOR of 1000 and Tumor Fraction of 3%:



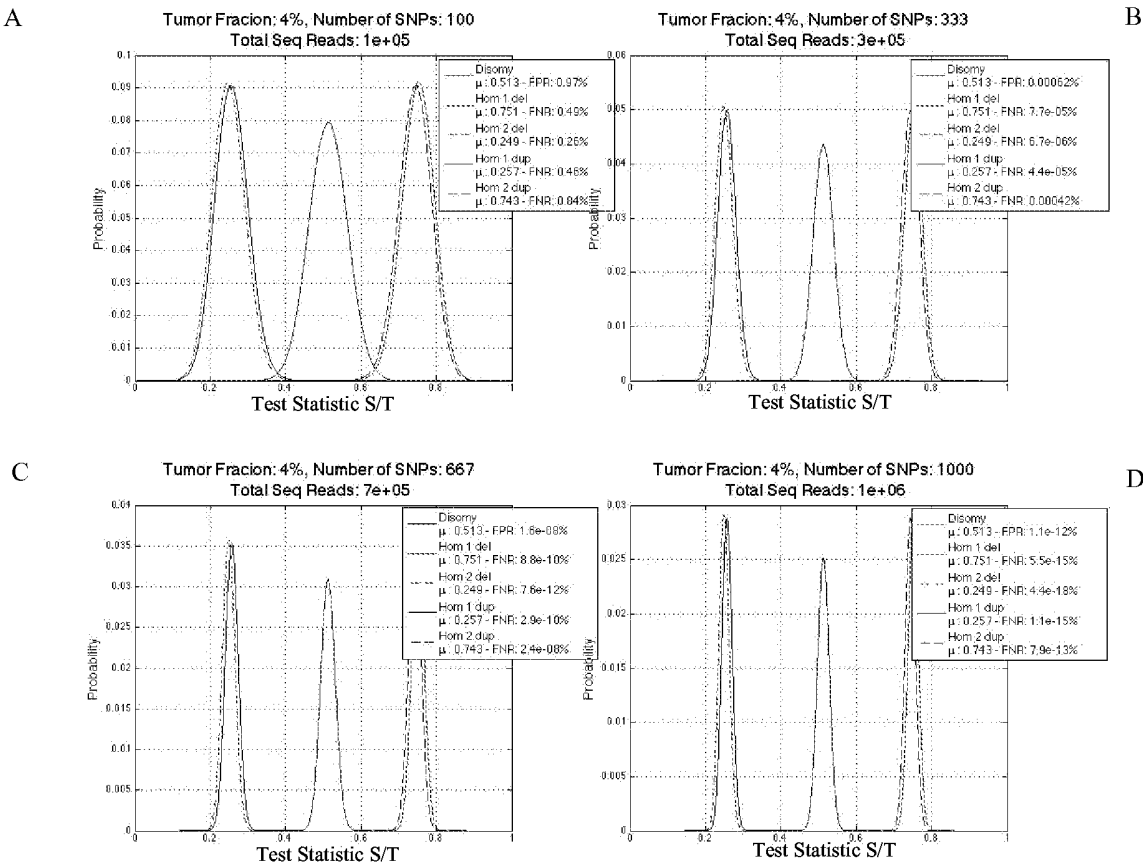
Figures 11A-11D

DOR of 1000 and Tumor Fraction of 4%:



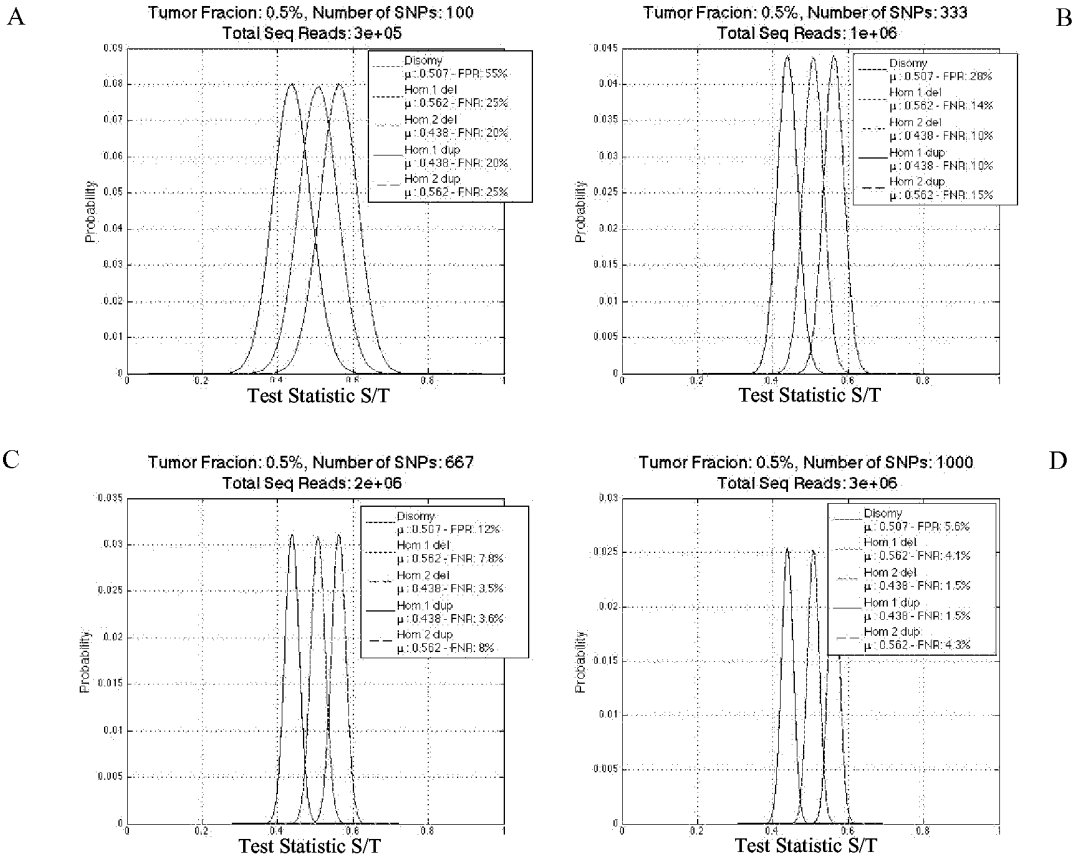
Figures 11A-11D

DOR of 1000 and Tumor Fraction of 4%:



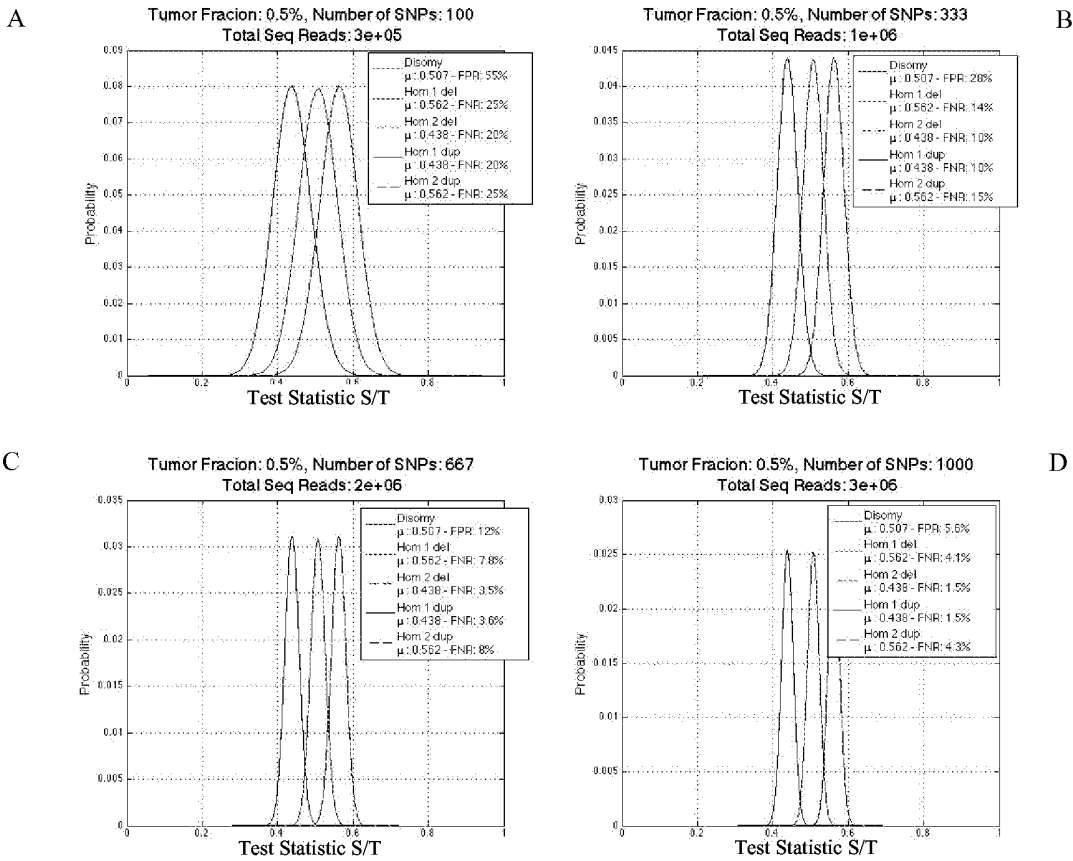
Figures 12A-12D

DOR of 3000 and Tumor Fraction of 0.5%:



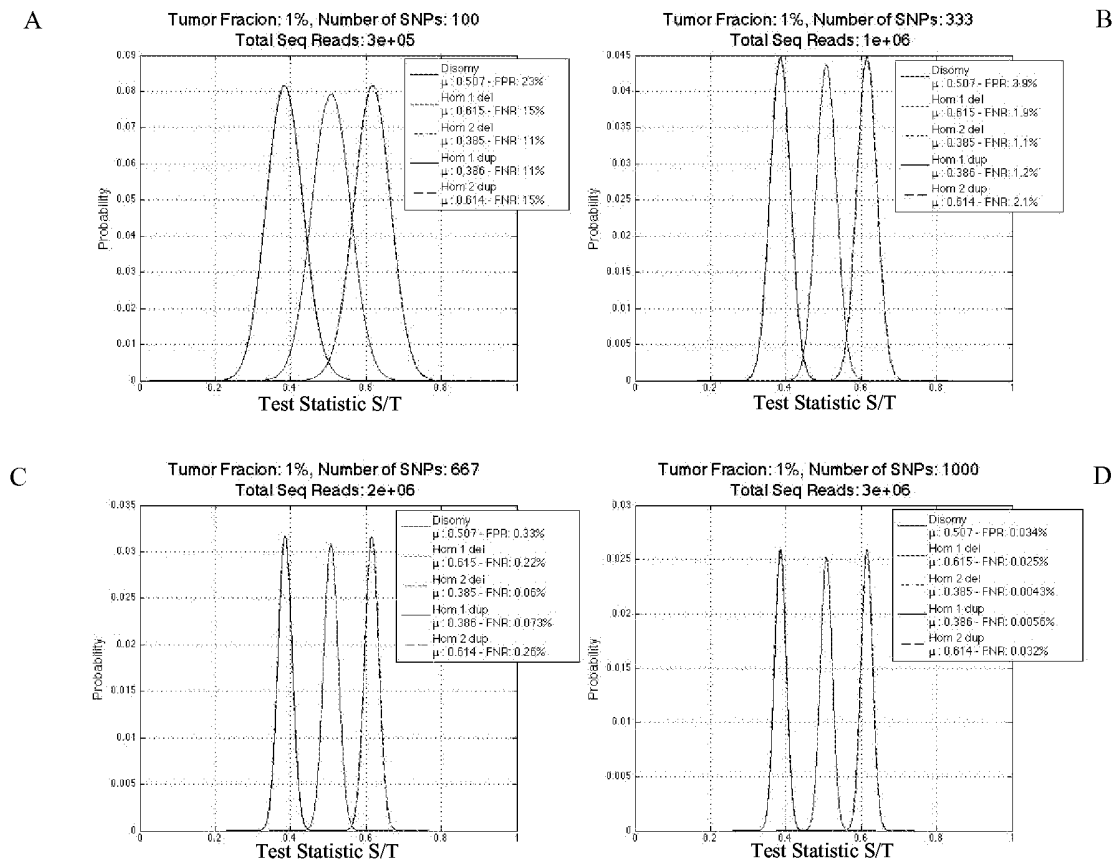
Figures 12A-12D

DOR of 3000 and Tumor Fraction of 0.5%:



Figures 13A-13D

DOR of 3000 and Tumor Fraction of 1%:



Figures 13A-13D

DOR of 3000 and Tumor Fraction of 1%:

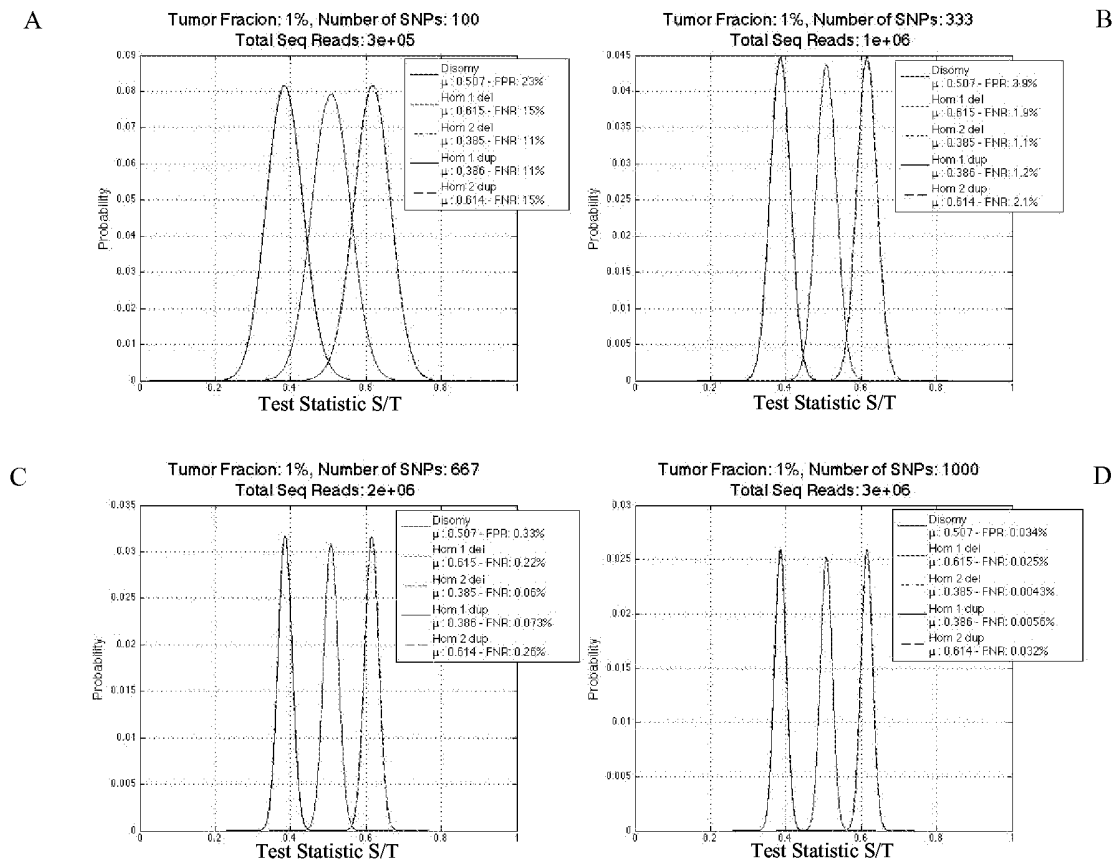


Figure 14

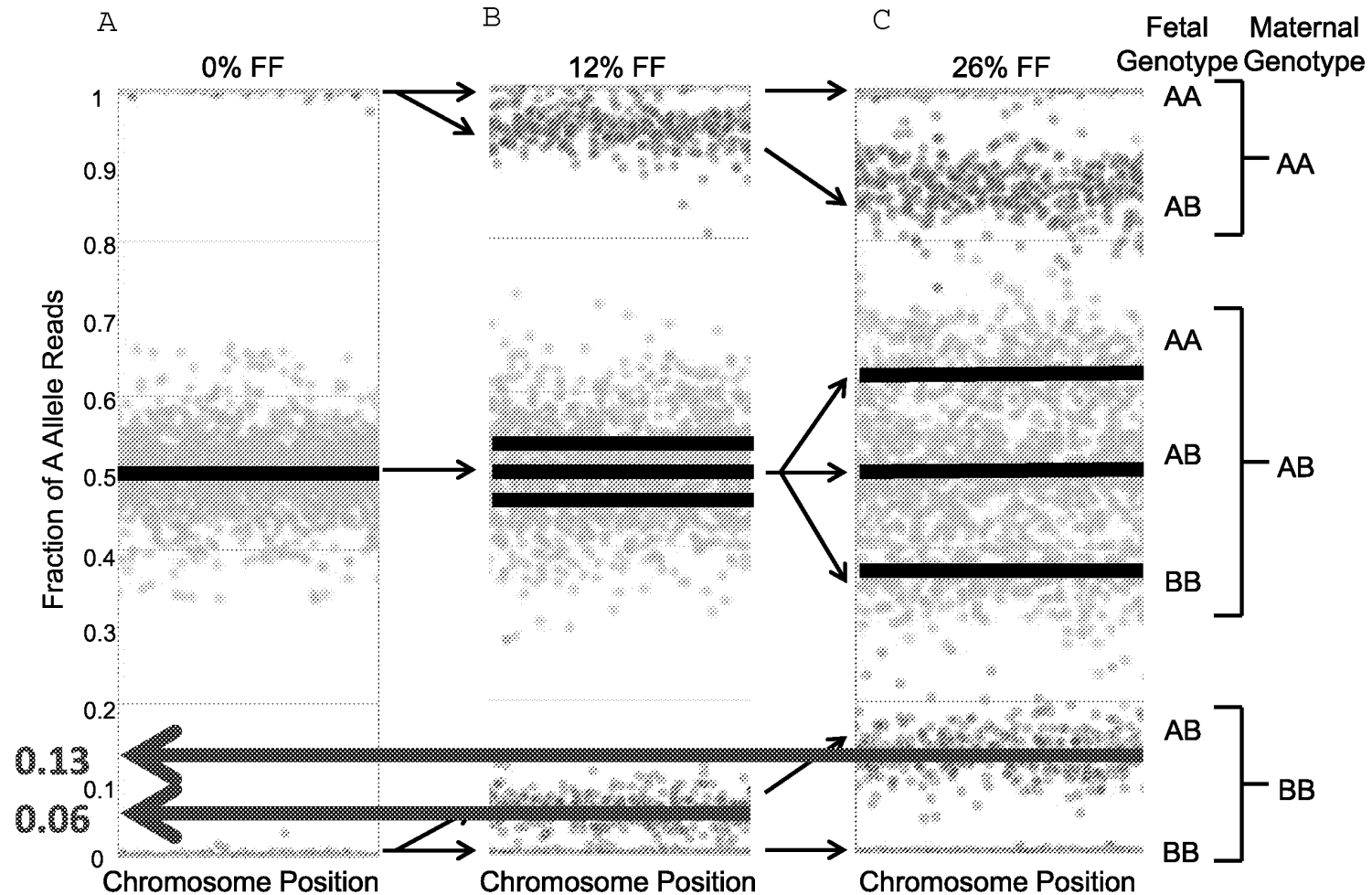
| <b>Syndrome</b>            | <b>Sensitivity</b>                   | <b>Specificity</b>                     |
|----------------------------|--------------------------------------|----------------------------------------|
| 22q11.2 Deletion/ DiGeorge | 97.8% (45/46)<br>[CI: 88.5 – 99.95%] | 99.2% (392/395)<br>[CI: 97.8 - 99.8%]  |
| Angelman                   | 100% (21/21)<br>[CI: 83.9 – 100%]    | 100% (418/418)<br>[CI: 99.1 – 100%]    |
| Cri du chat                | 100% (24/24)<br>[CI: 85.7 – 100%]    | 99.8% (415/416)<br>[CI: 98.7 – 99.99%] |
| Monosomy 1p36              | 100% (1/1)<br>[CI: 2.5 – 100%]       | 100% (438/438)<br>[CI: 99.2 – 100%]    |
| Prader-Willi               | 100% (15/15)<br>[CI: 78.2 – 100%]    | 100% (424/424)<br>[CI: 99.1 – 100%]    |
| Wolf-Hirschhorn            | 100% (2/2)<br>[CI: 15.8 – 100%]      | 99.8% (437/438)<br>[CI: 98.7 – 100%]   |

Figure 14

| <b>Syndrome</b>            | <b>Sensitivity</b>                   | <b>Specificity</b>                     |
|----------------------------|--------------------------------------|----------------------------------------|
| 22q11.2 Deletion/ DiGeorge | 97.8% (45/46)<br>[CI: 88.5 – 99.95%] | 99.2% (392/395)<br>[CI: 97.8 - 99.8%]  |
| Angelman                   | 100% (21/21)<br>[CI: 83.9 – 100%]    | 100% (418/418)<br>[CI: 99.1 – 100%]    |
| Cri du chat                | 100% (24/24)<br>[CI: 85.7 – 100%]    | 99.8% (415/416)<br>[CI: 98.7 – 99.99%] |
| Monosomy 1p36              | 100% (1/1)<br>[CI: 2.5 – 100%]       | 100% (438/438)<br>[CI: 99.2 – 100%]    |
| Prader-Willi               | 100% (15/15)<br>[CI: 78.2 – 100%]    | 100% (424/424)<br>[CI: 99.1 – 100%]    |
| Wolf-Hirschhorn            | 100% (2/2)<br>[CI: 15.8 – 100%]      | 99.8% (437/438)<br>[CI: 98.7 – 100%]   |

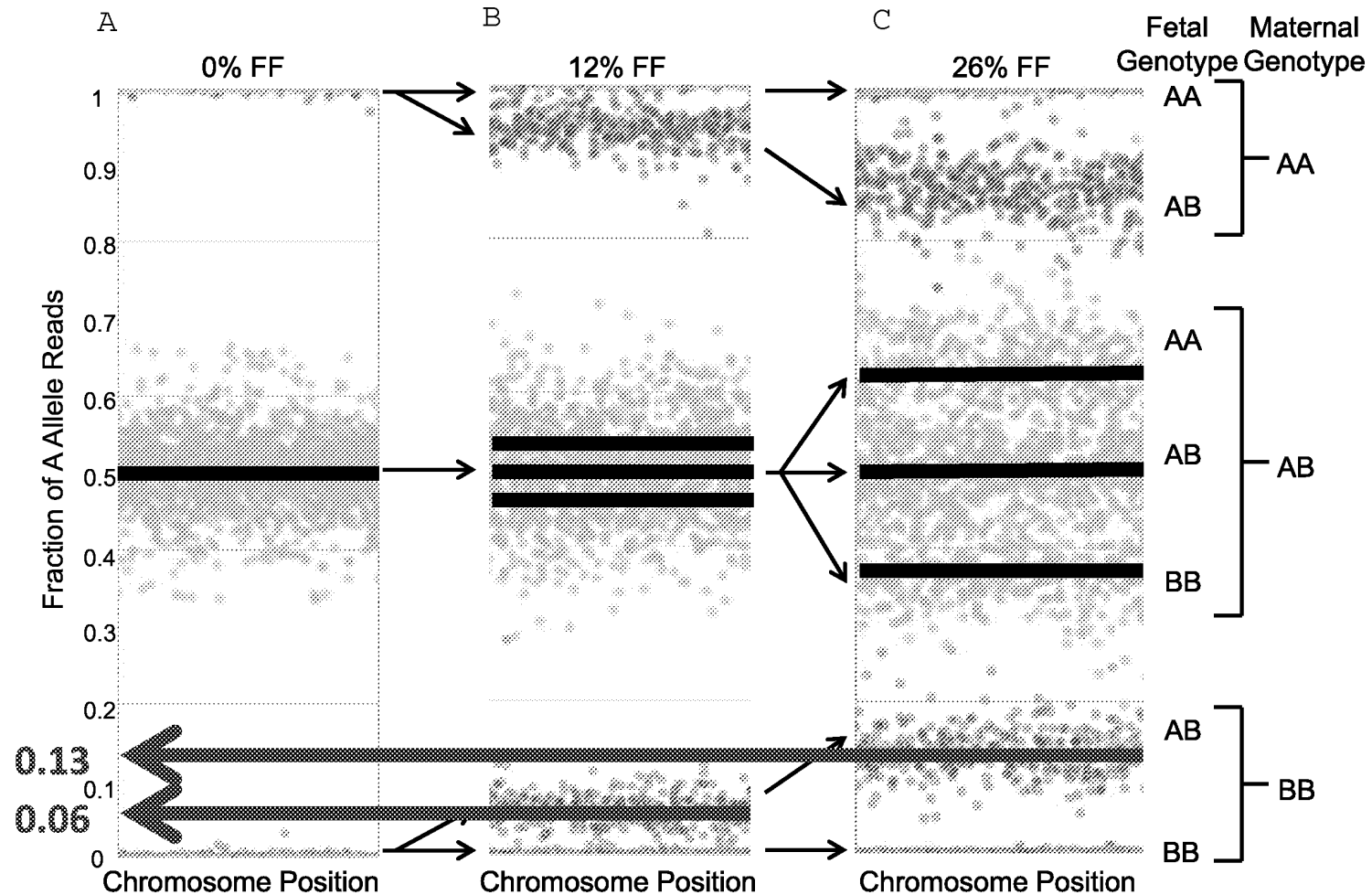
Figures 15A-15C

# Allele Frequency Plots: Visualizing Data

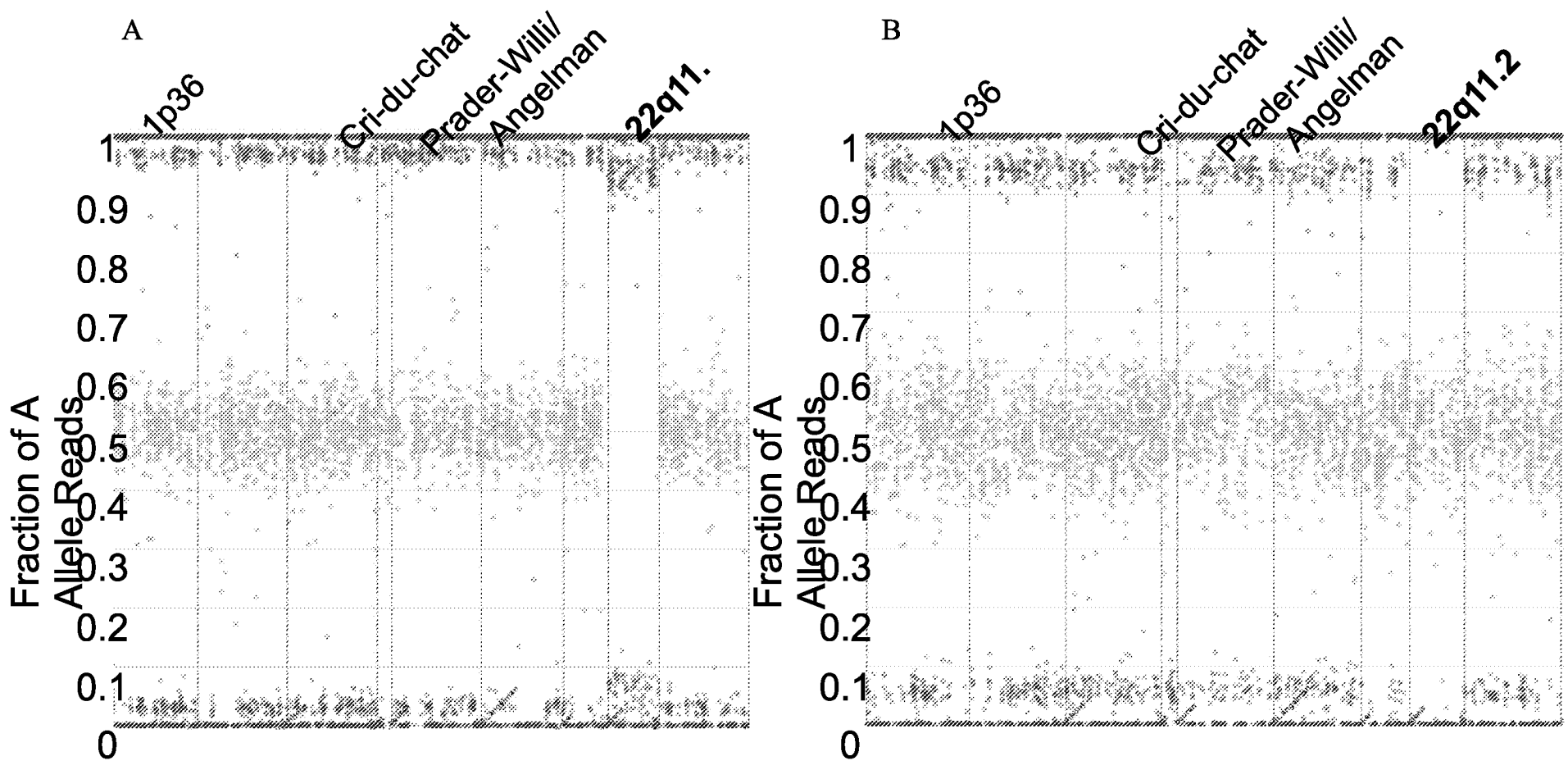


Figures 15A-15C

# Allele Frequency Plots: Visualizing Data



Figures 16A and 16B



Figures 16A and 16B

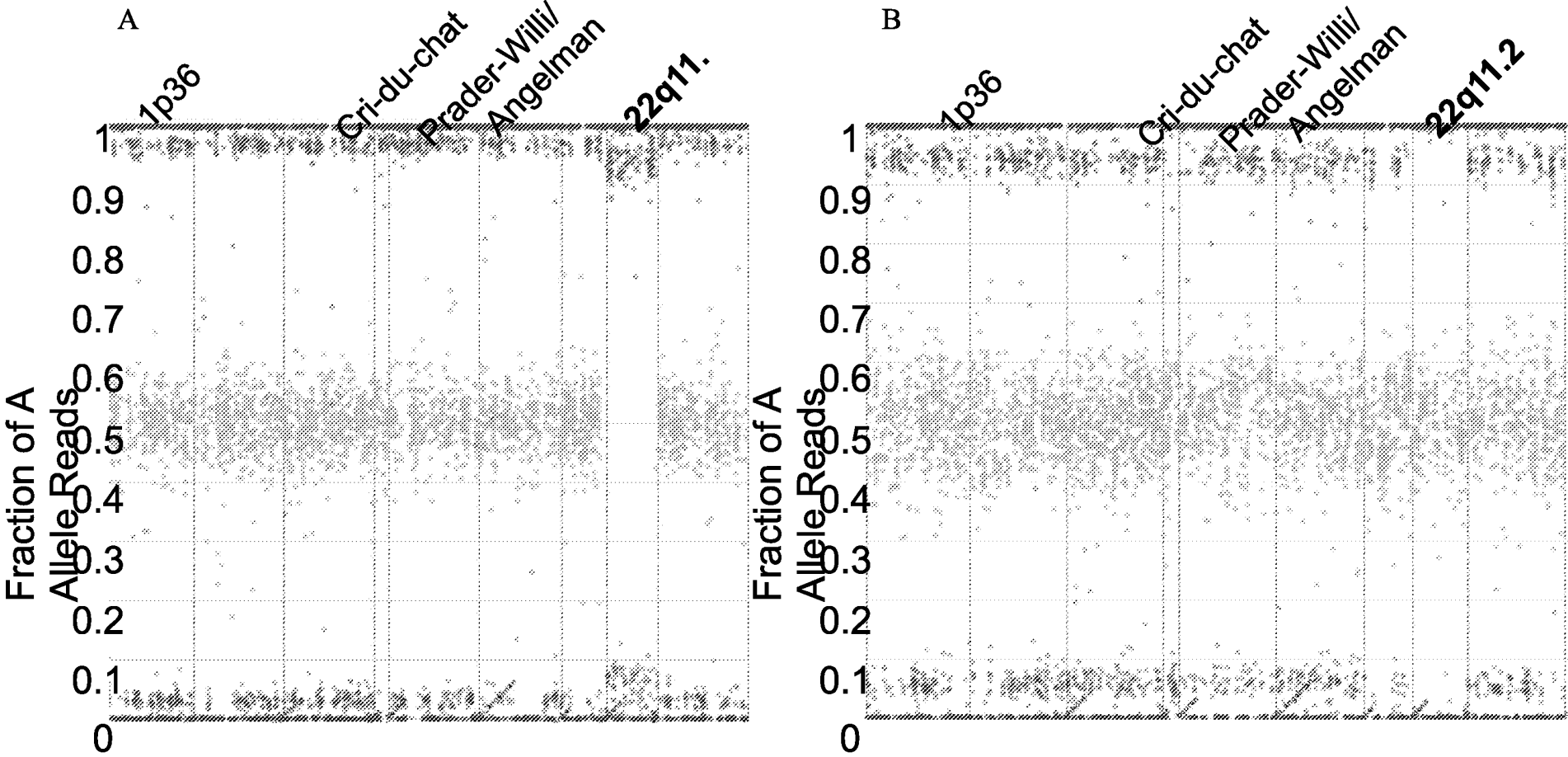


Figure 17

## Maternally Inherited Cri-du-Chat Syndrome

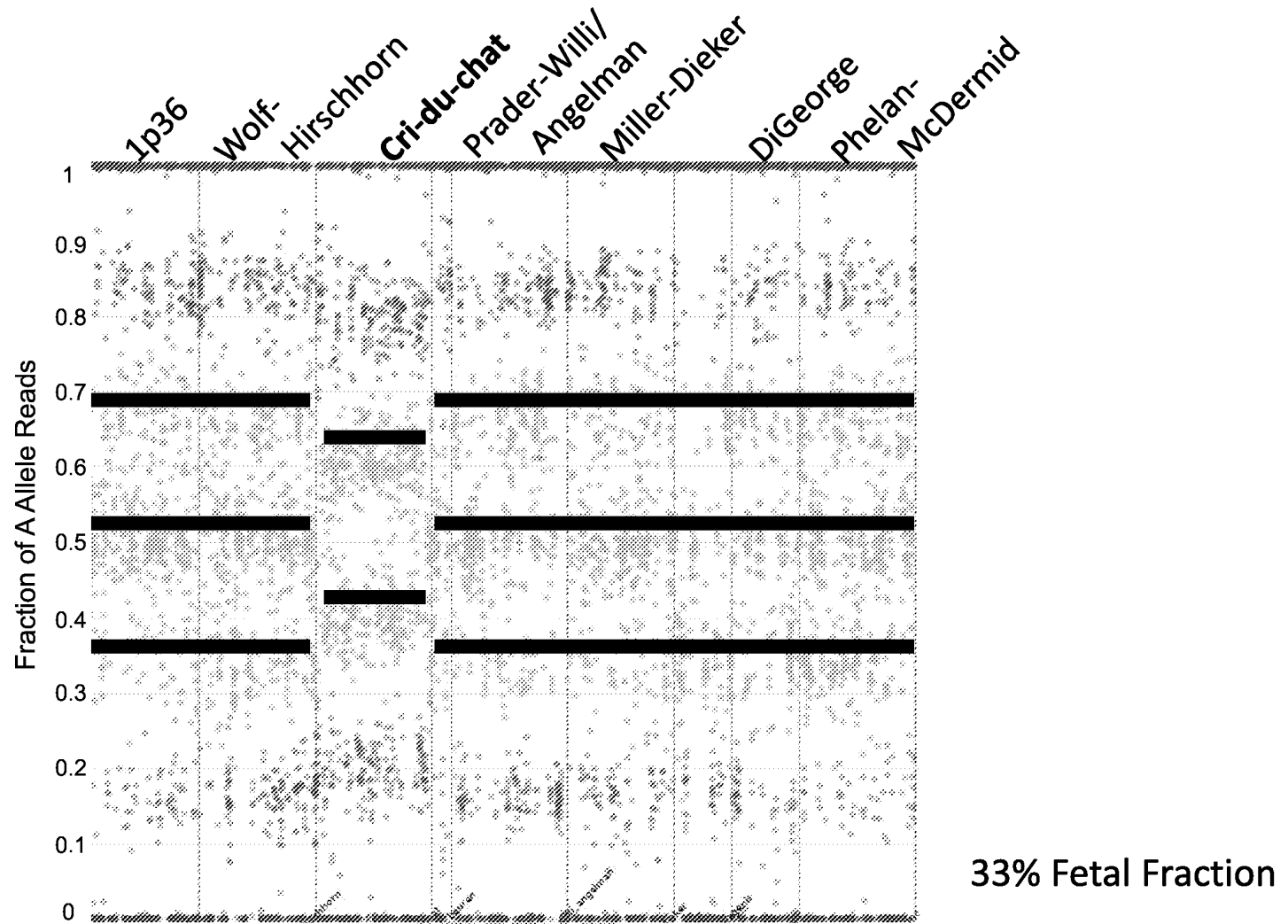


Figure 17

## Maternally Inherited Cri-du-Chat Syndrome

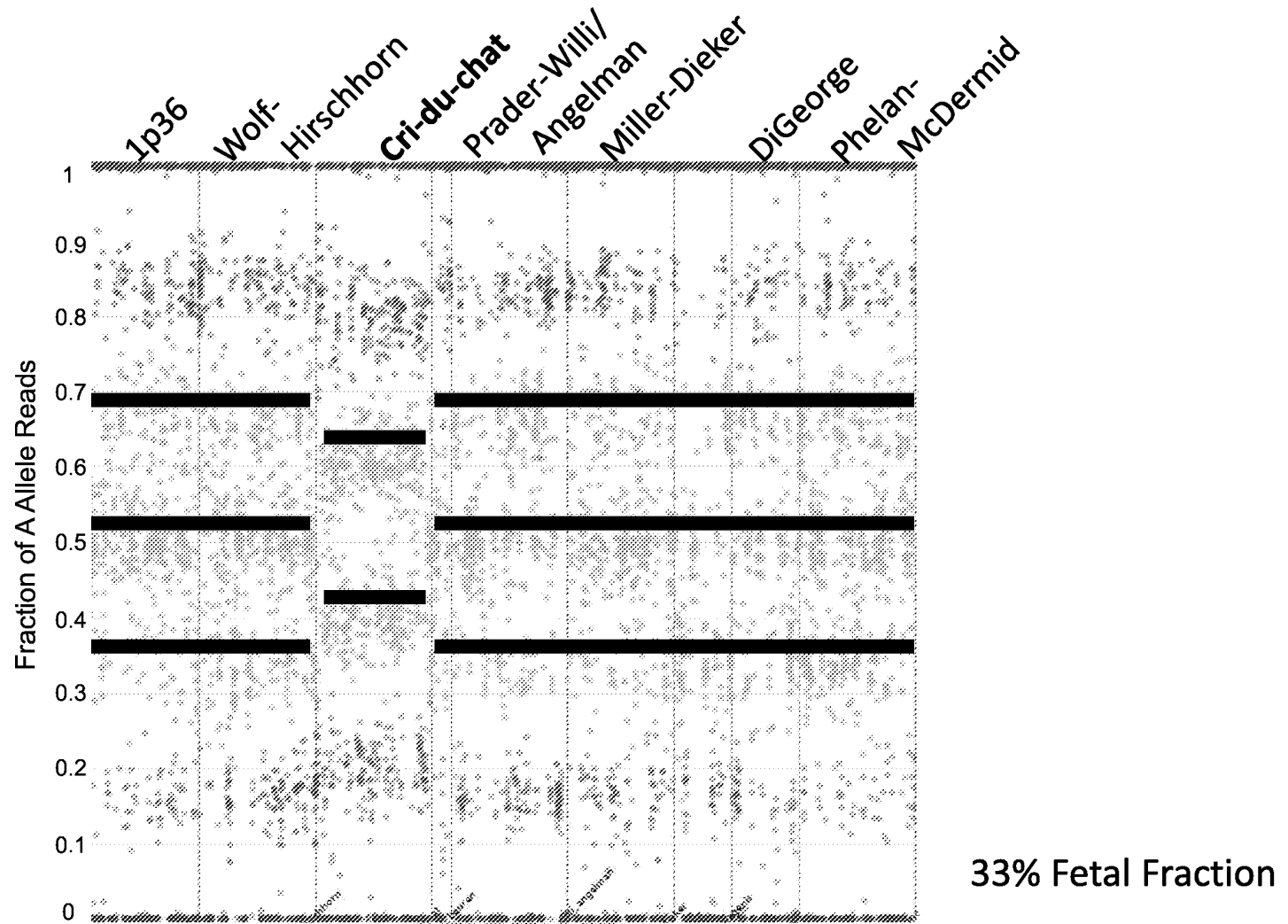


Figure 18

# Paternally Inherited Wolf-Hirschhorn Syndrome

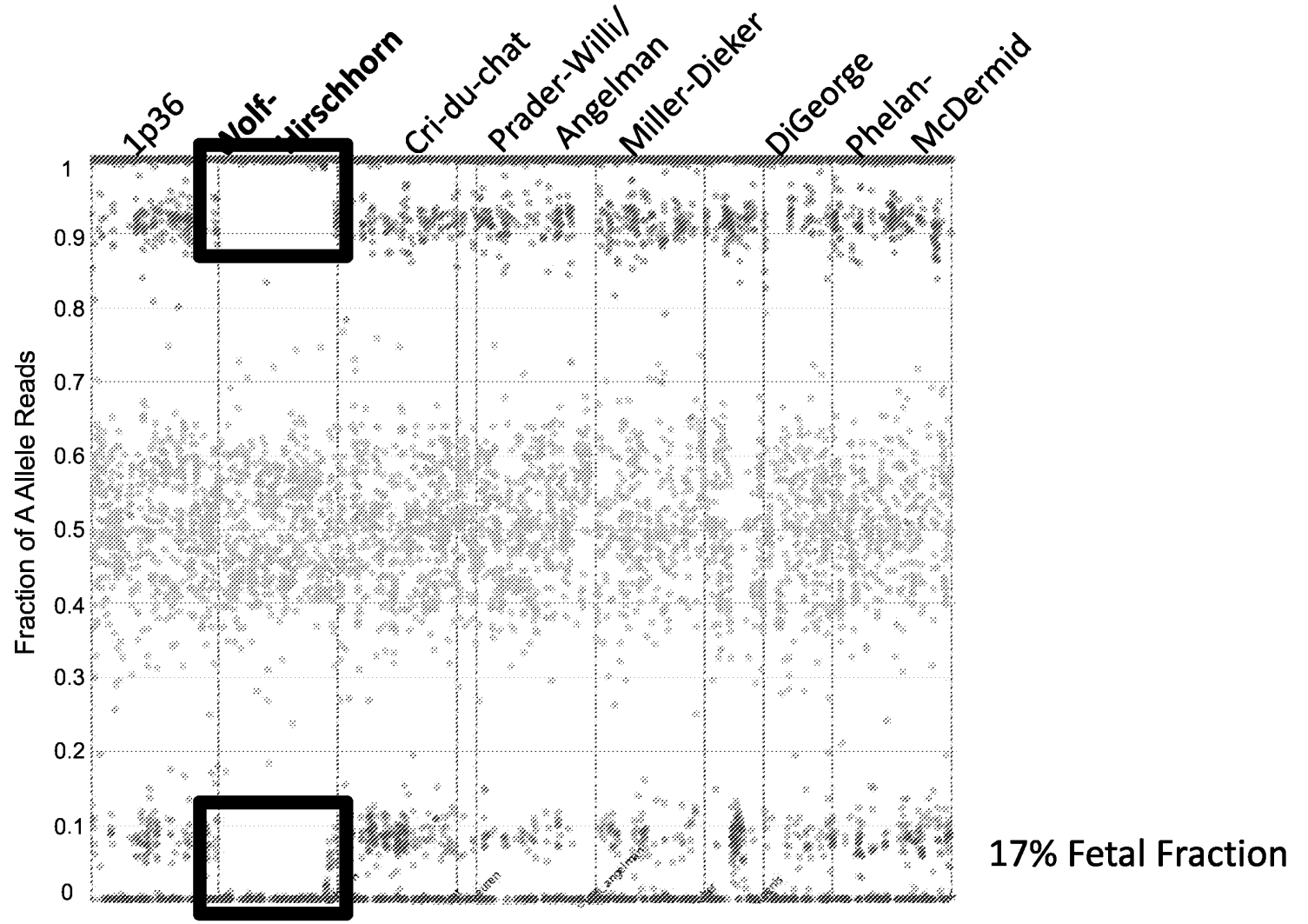


Figure 18

# Paternally Inherited Wolf-Hirschhorn Syndrome

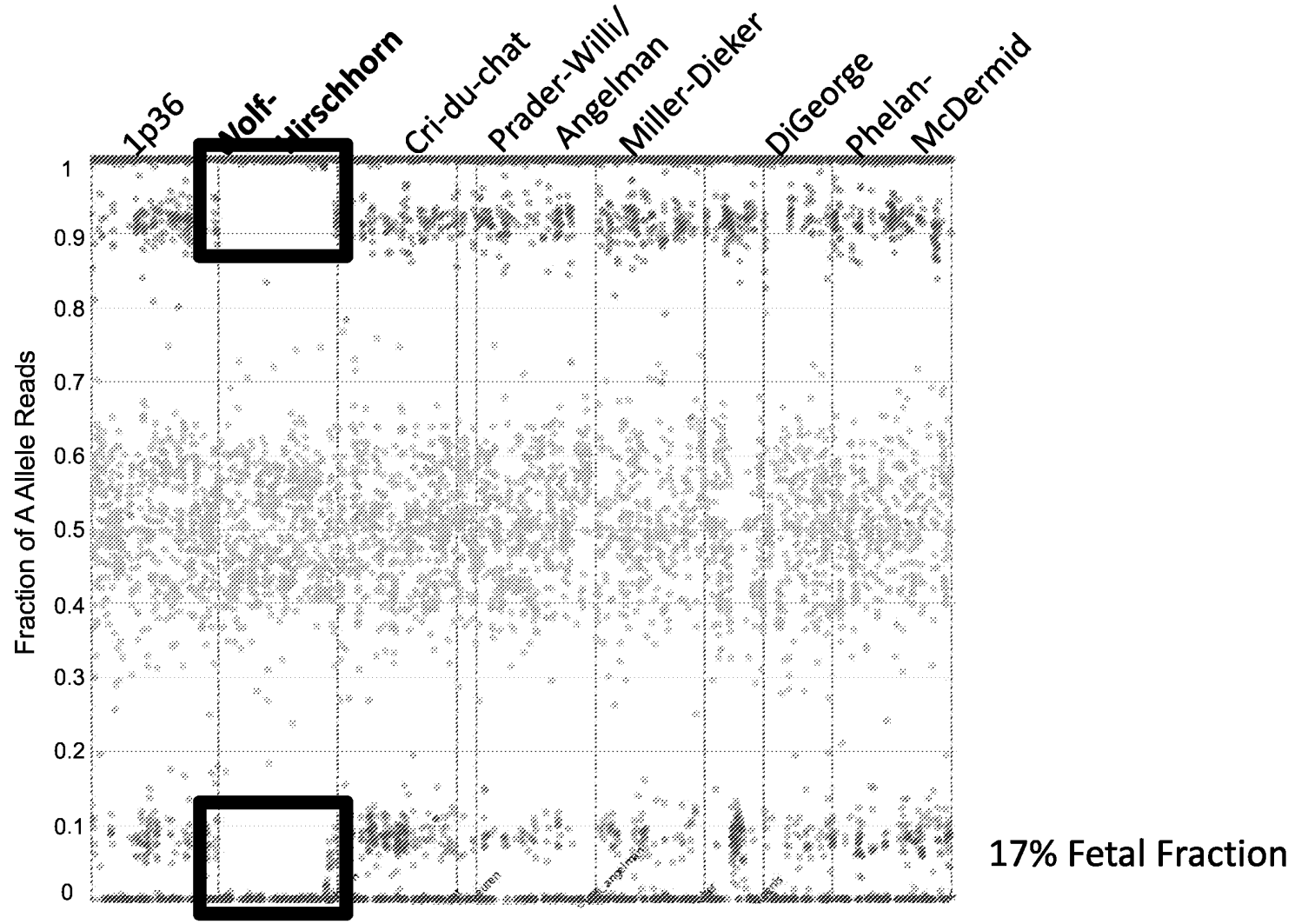


FIG. 19A 16% Father in Daughter

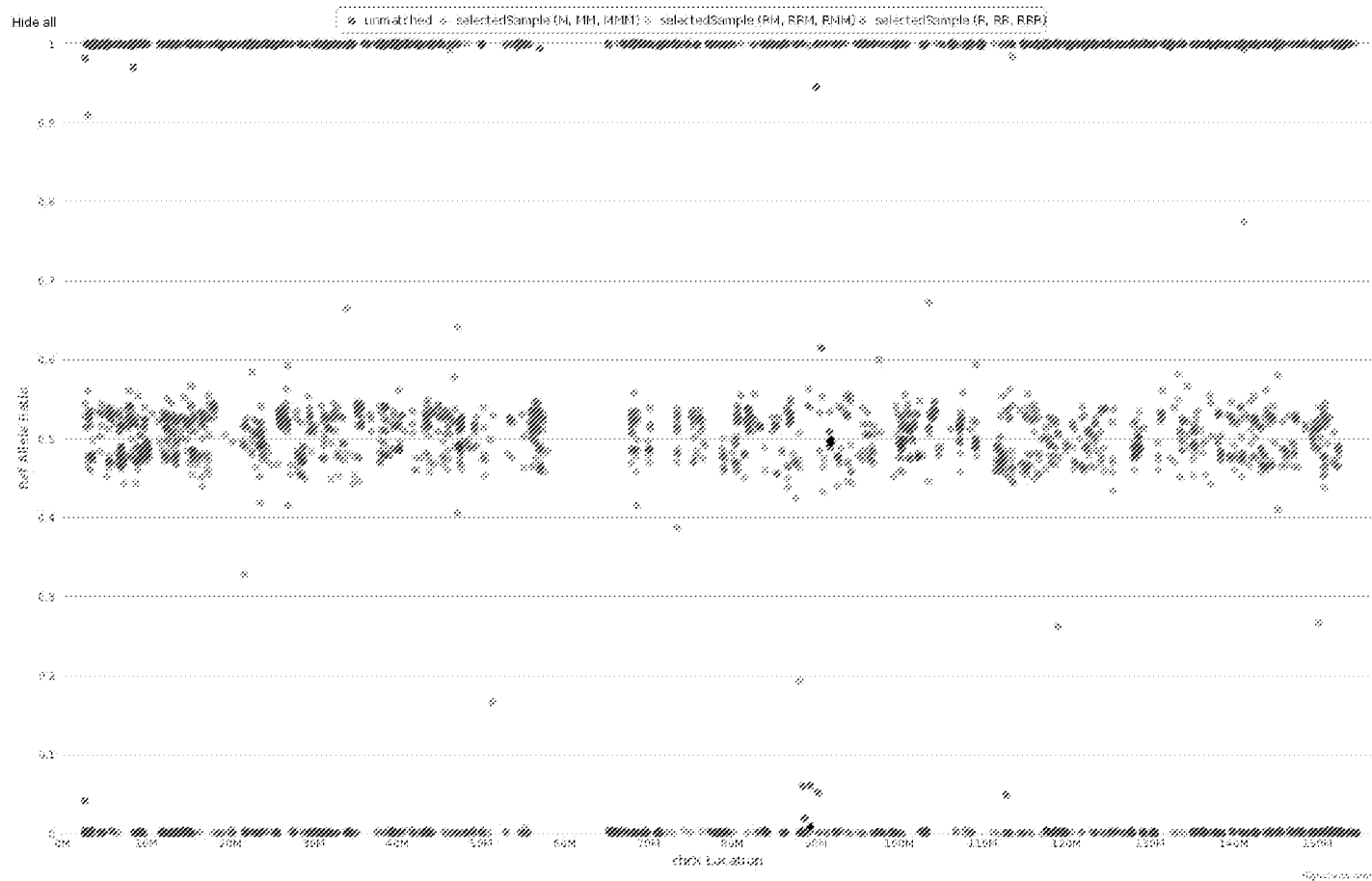


FIG. 19A 16% Father in Daughter

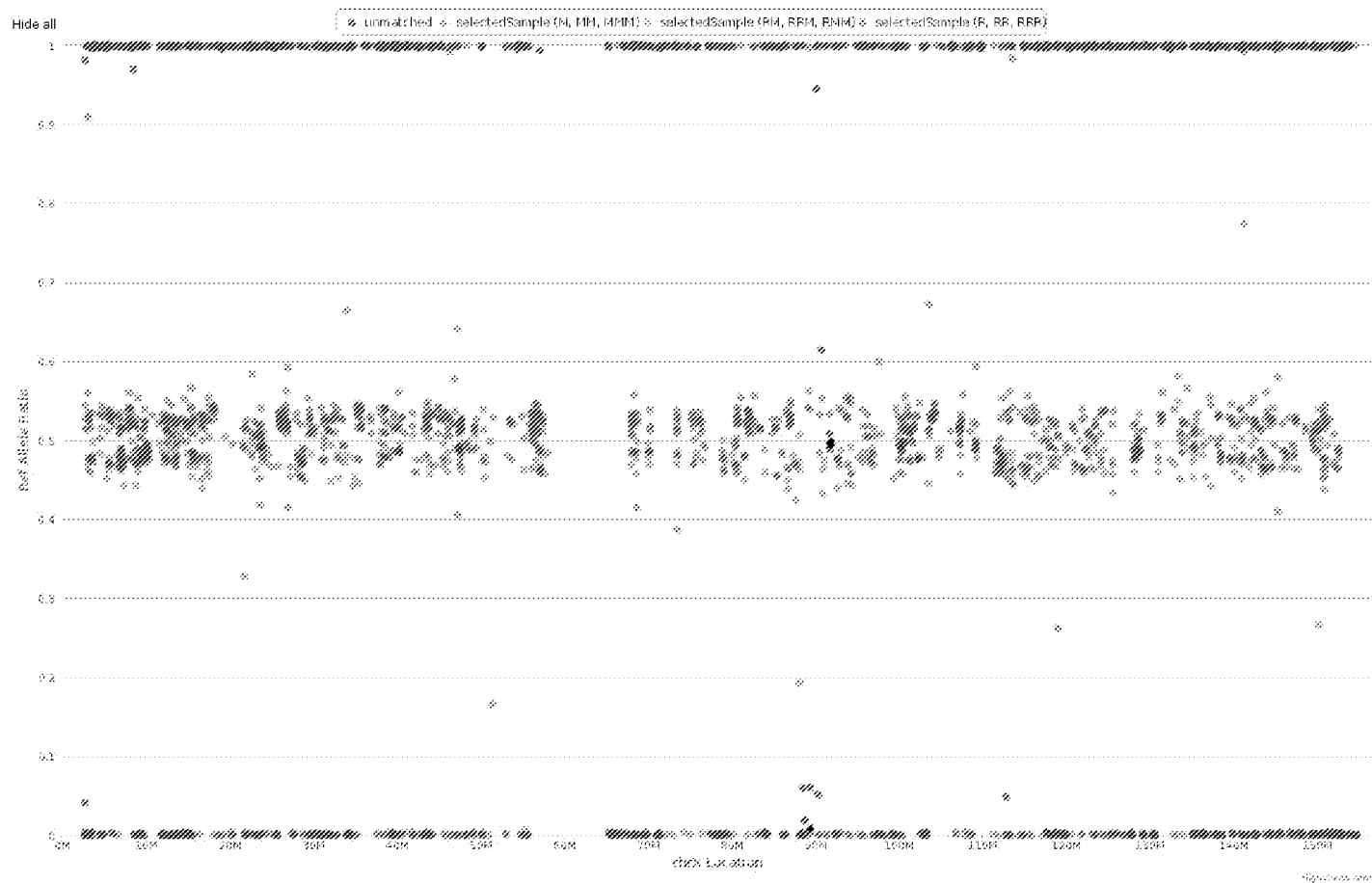


FIG. 19B 10% Father in Daughter

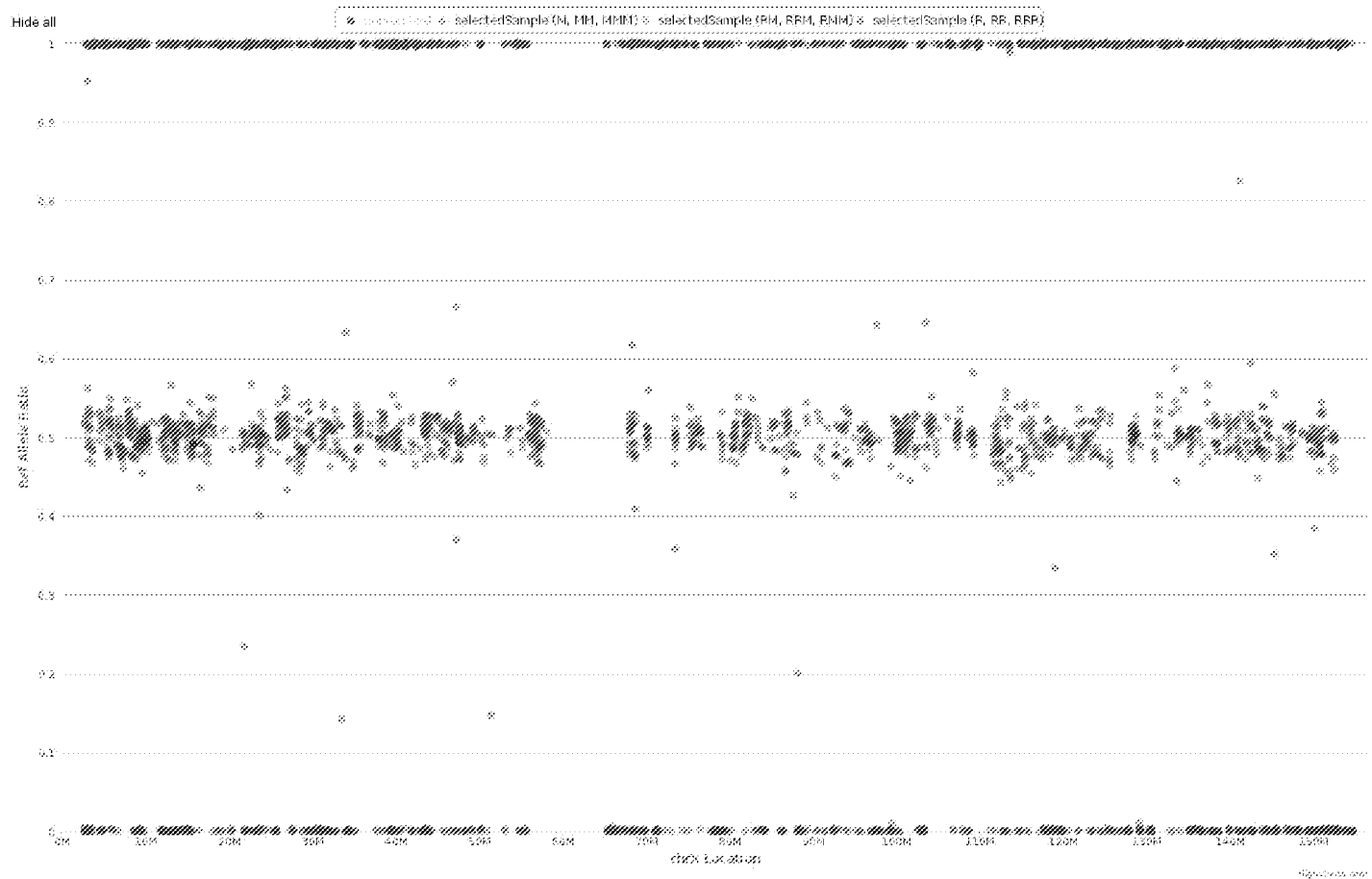


FIG. 19B 10% Father in Daughter

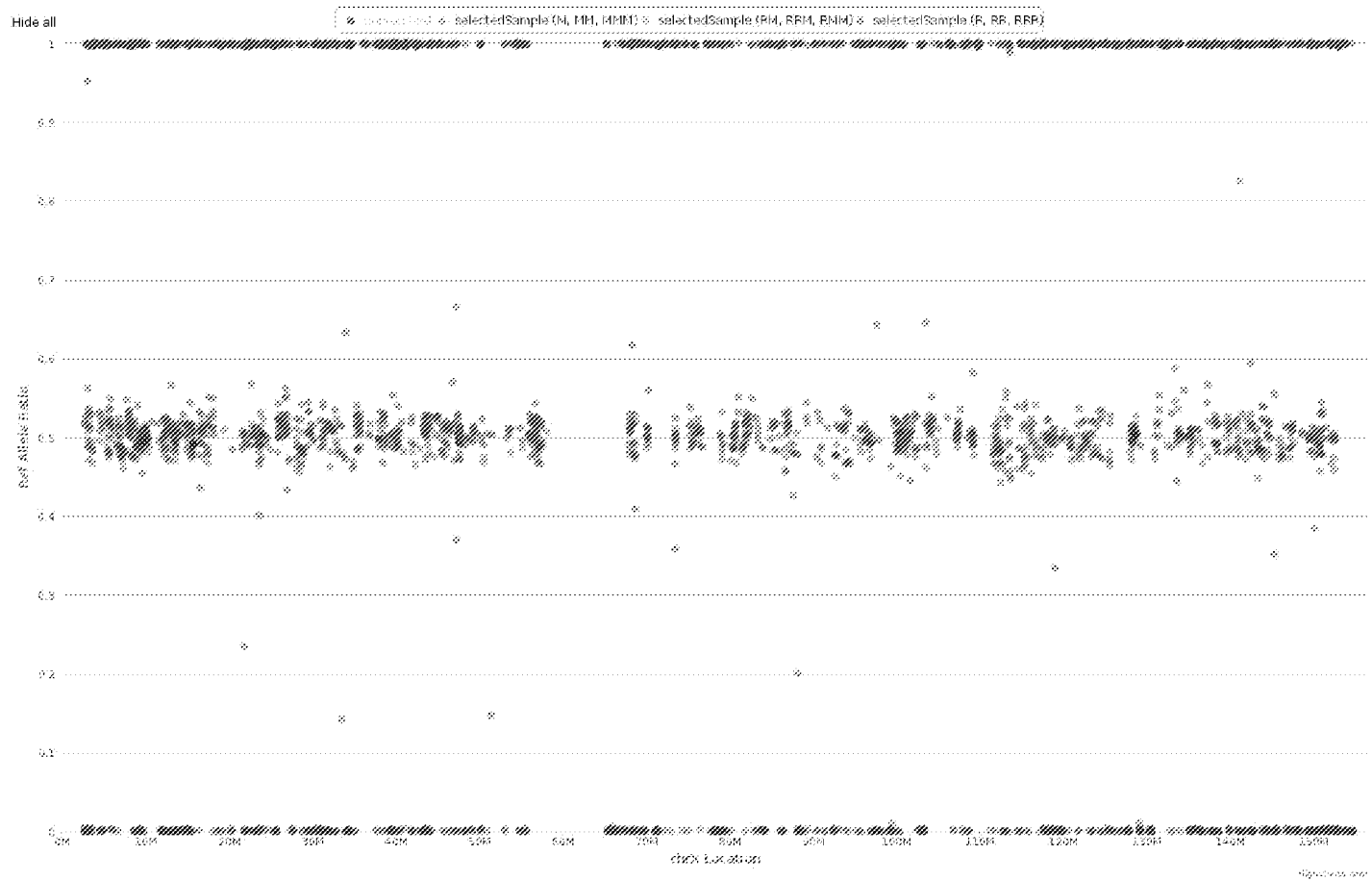


FIG. 19C 1% Father in Daughter

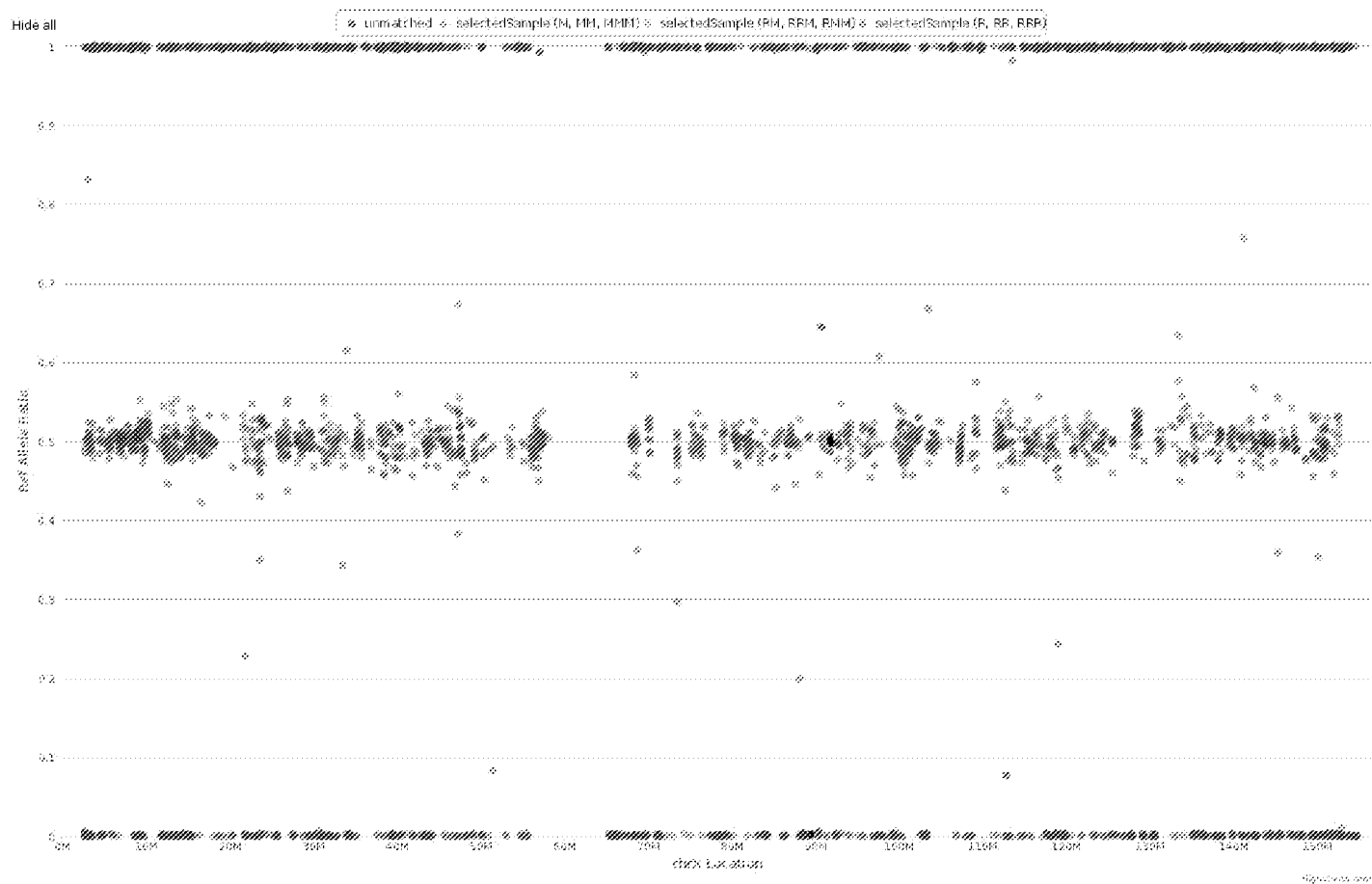


FIG. 19C 1% Father in Daughter

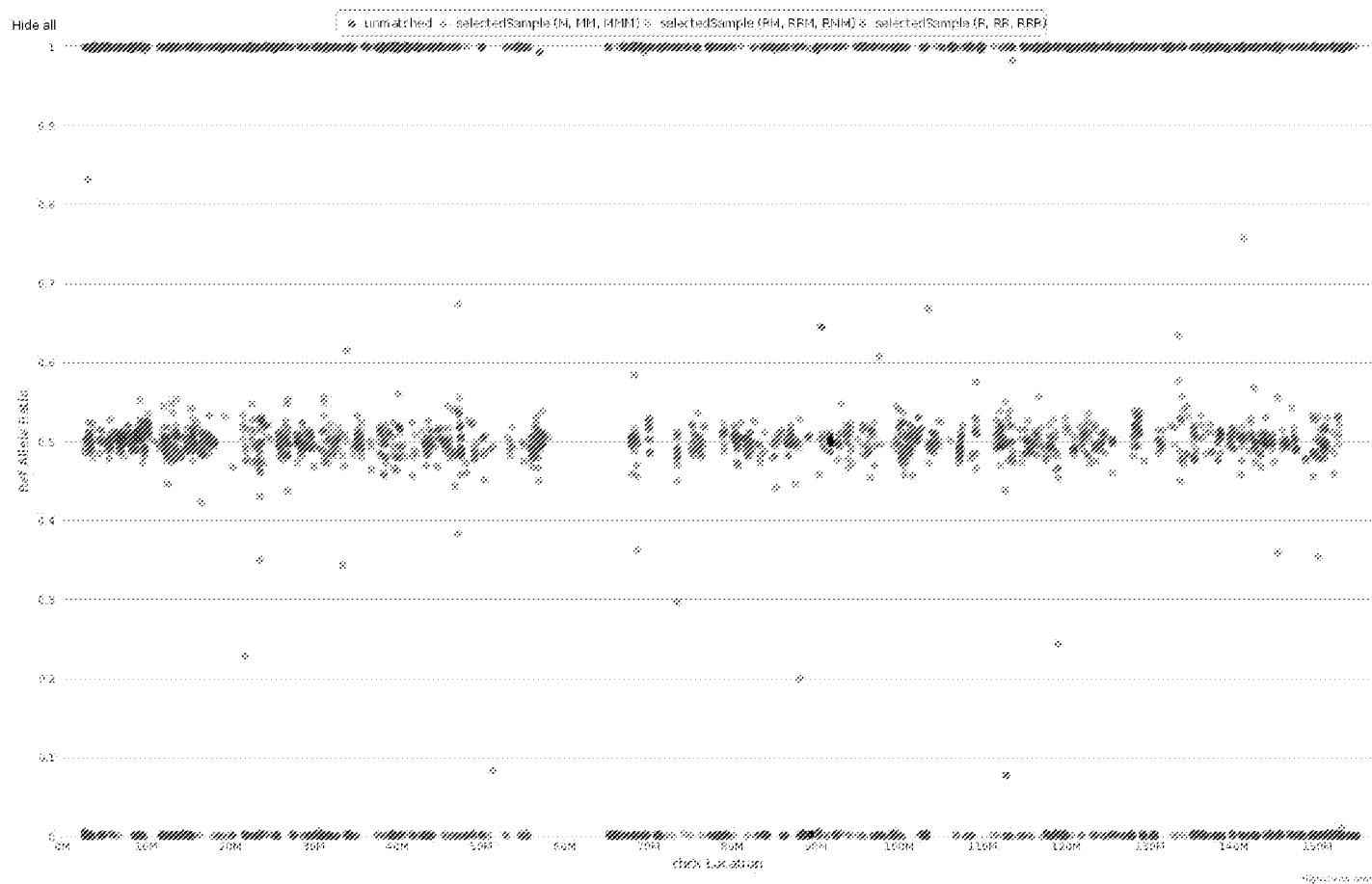


FIG. 19D 0.1% Father in Daughter

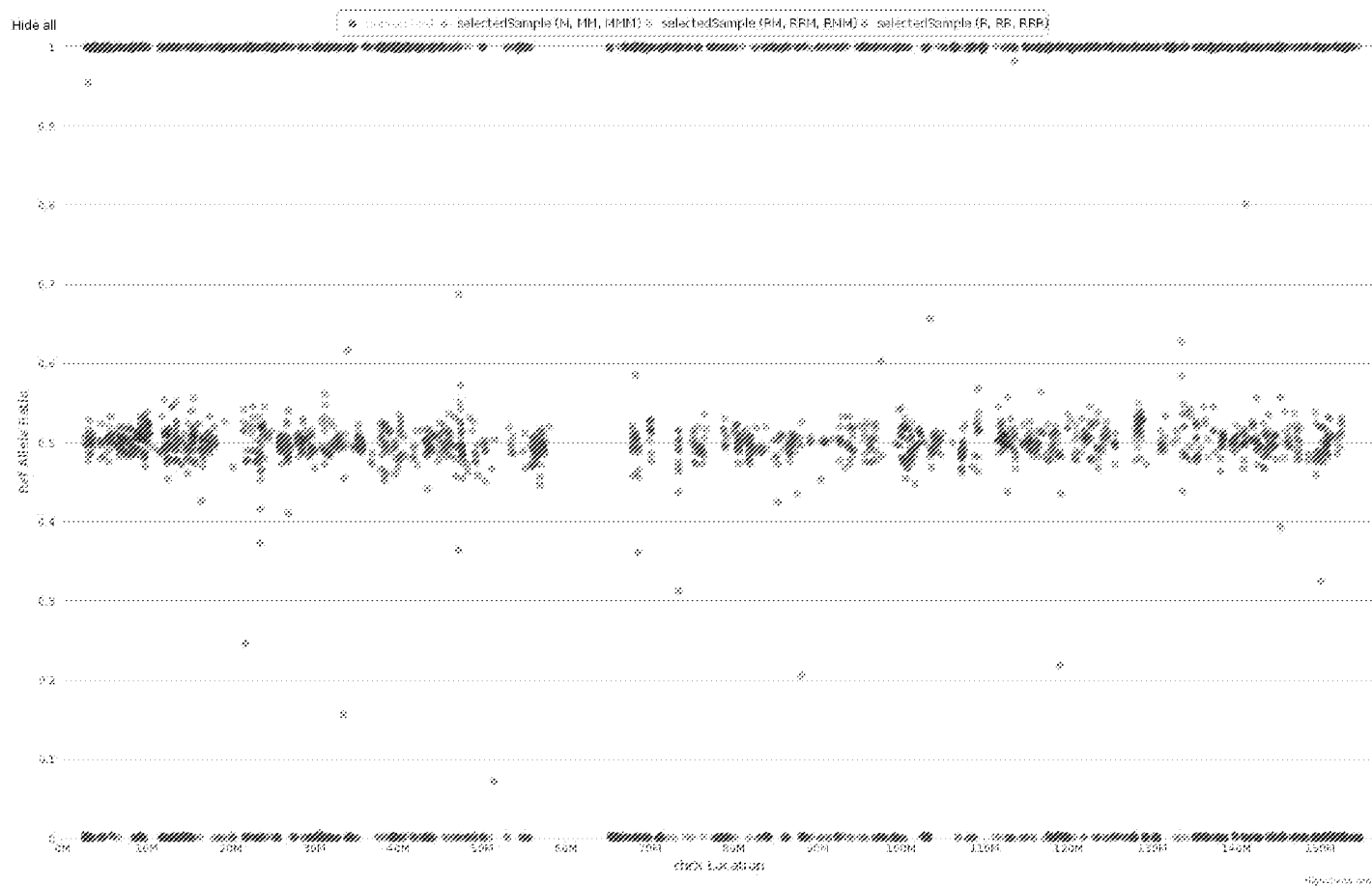
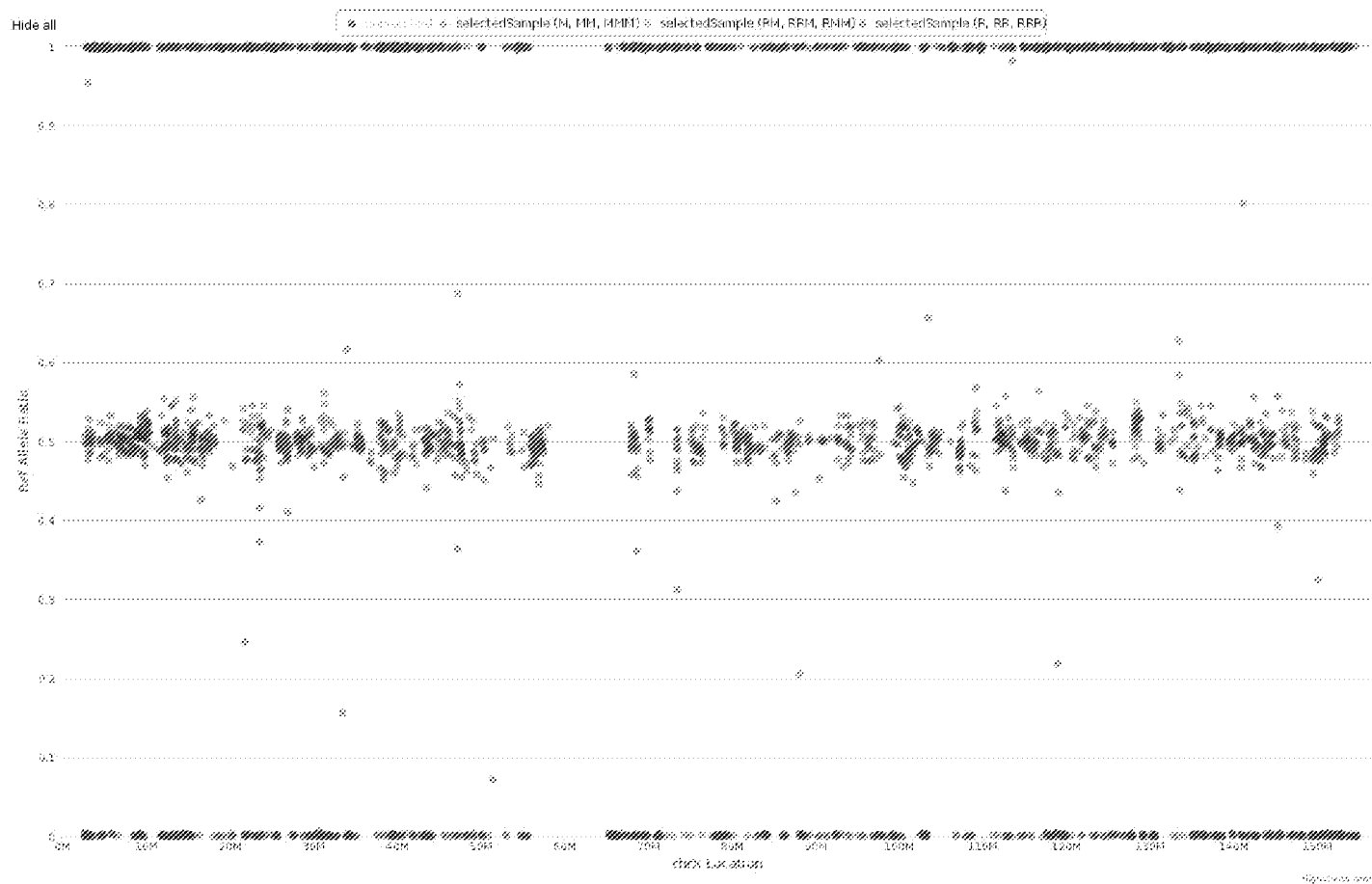
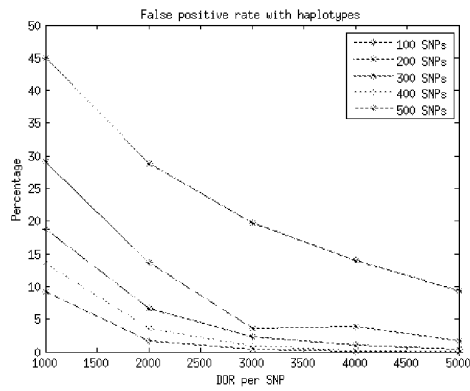


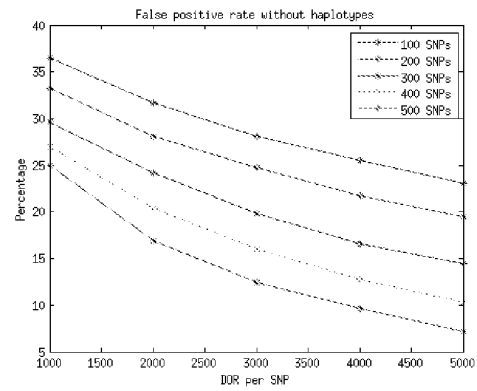
FIG. 19D 0.1% Father in Daughter



Figures 20A and 20B

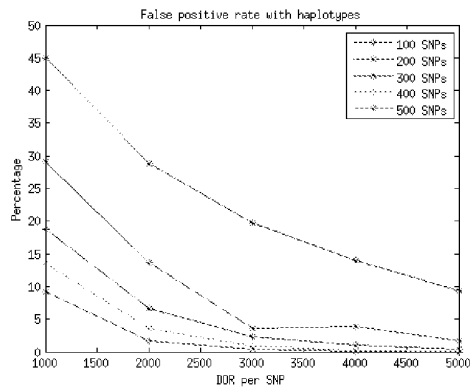


(a) With Haplotypes

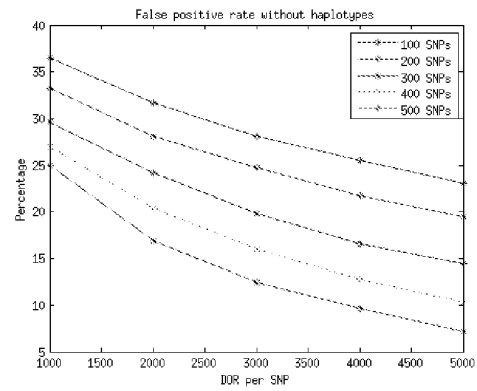


(b) Without Haplotypes

Figures 20A and 20B

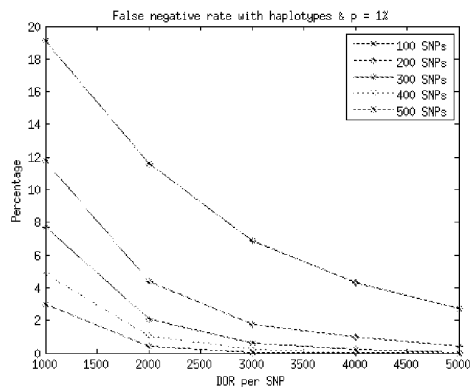


(a) With Haplotypes

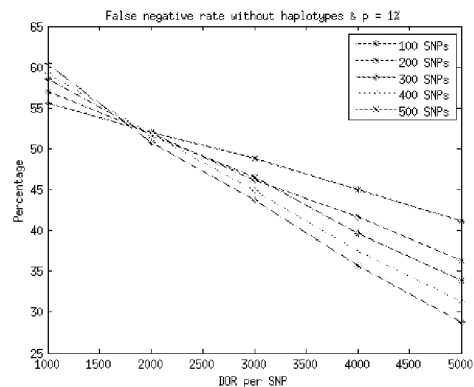


(b) Without Haplotypes

Figures 21A and 21B

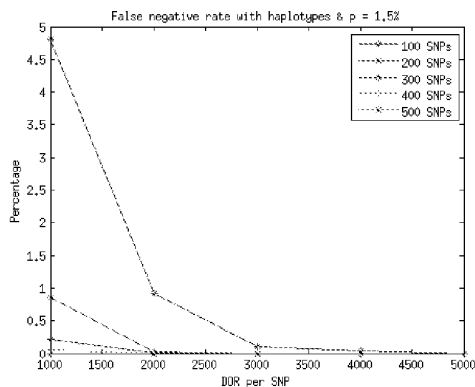


(a) With Haplotypes

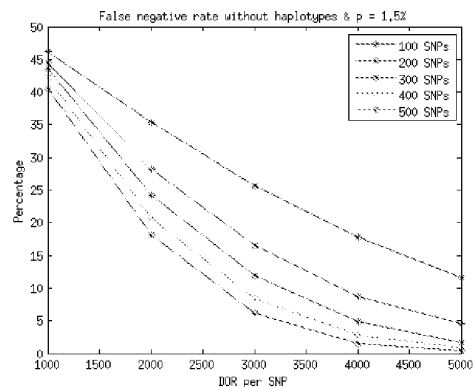


(b) Without Haplotypes

Figures 22A and 22B

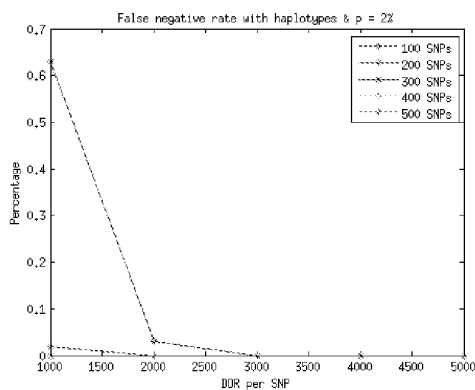


(a) With Haplotypes

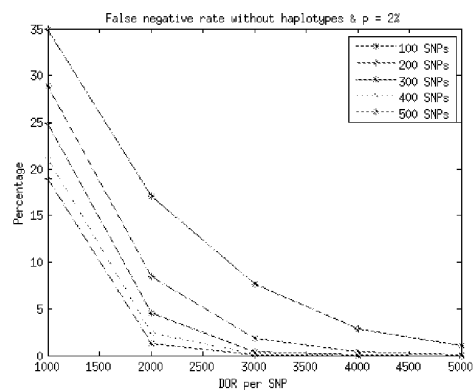


(b) Without Haplotypes

Figures 23A and 23B

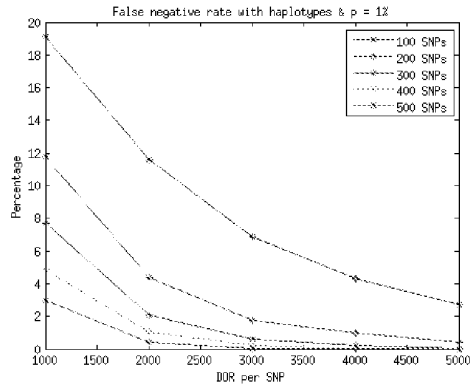


(a) With Haplotypes

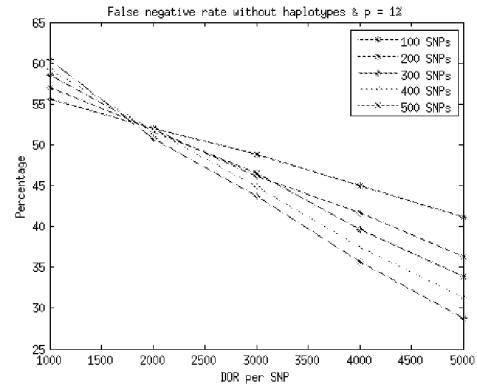


(b) Without Haplotypes

Figures 21A and 21B

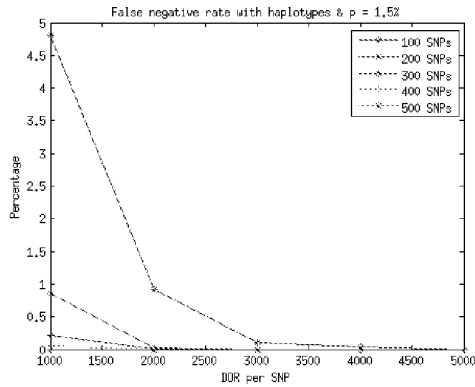


(a) With Haplotypes

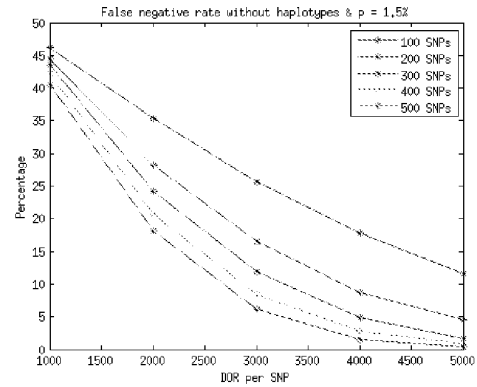


(b) Without Haplotypes

Figures 22A and 22B

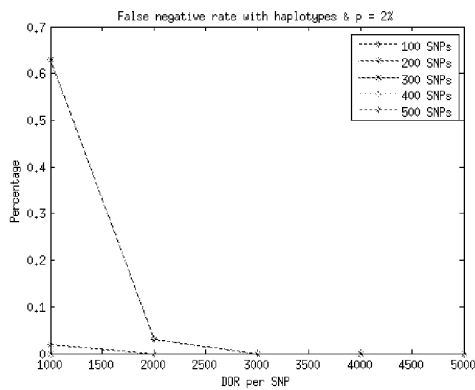


(a) With Haplotypes

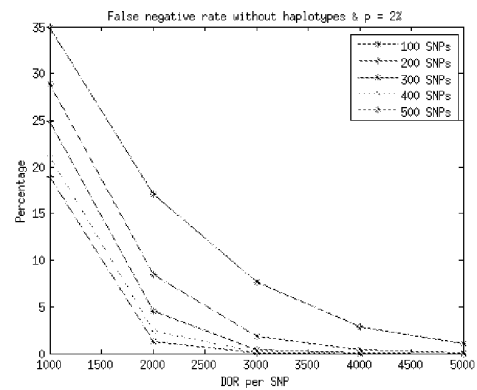


(b) Without Haplotypes

Figures 23A and 23B

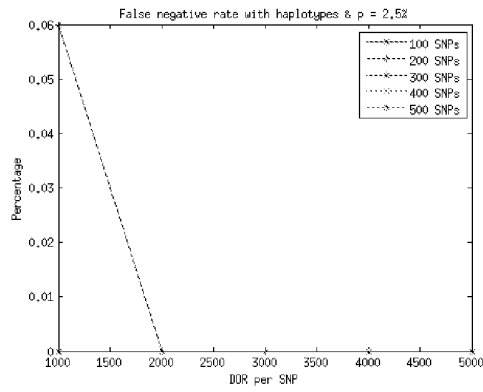


(a) With Haplotypes

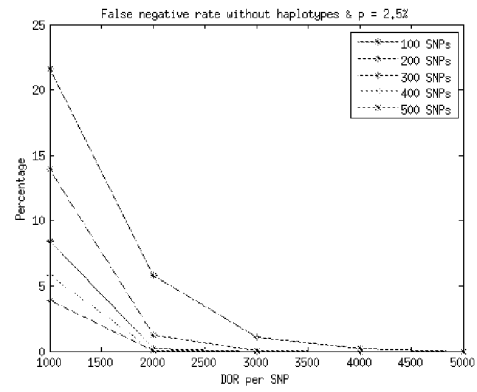


(b) Without Haplotypes

Figures 24A and 24B

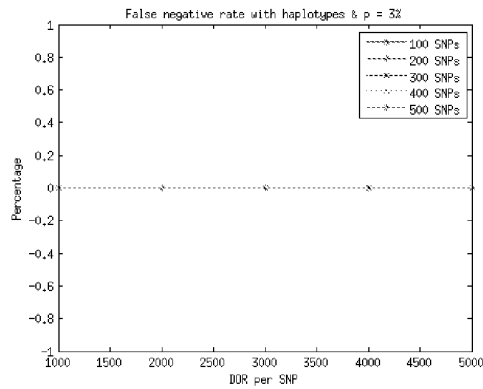


(a) With Haplotypes

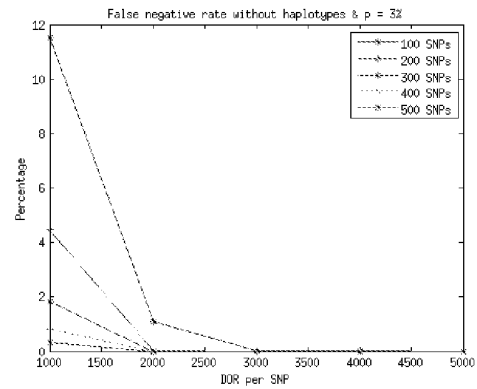


(b) Without Haplotypes

Figures 25A and 25B

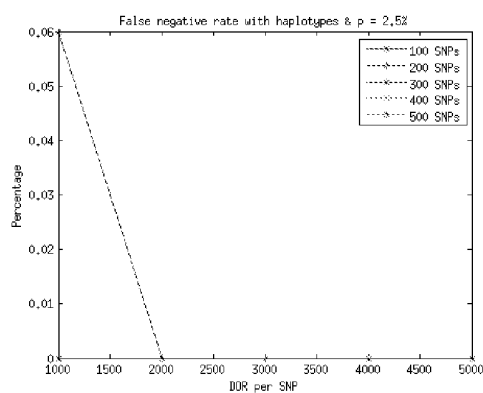


(a) With Haplotypes

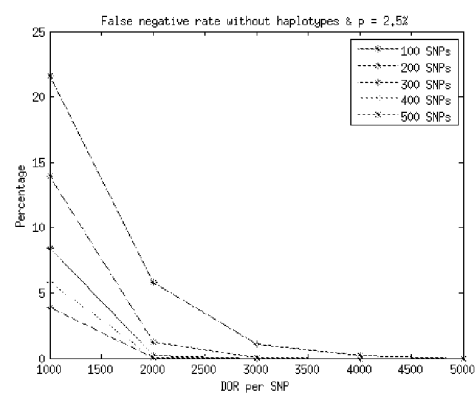


(b) Without Haplotypes

Figures 24A and 24B

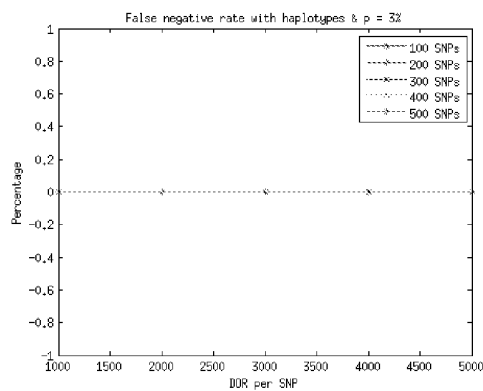


(a) With Haplotypes

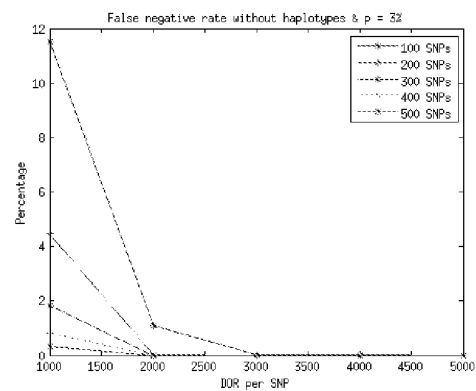


(b) Without Haplotypes

Figures 25A and 25B



(a) With Haplotypes



(b) Without Haplotypes

Figure 26

|               | S = 1000, D = 1000 | S = 500, D = 1000 | S = 1000, D = 500 | S = 500, D = 500 |
|---------------|--------------------|-------------------|-------------------|------------------|
| With Phase    | 1.50%              | 8.80%             | 10.20%            | 22.30%           |
| Without Phase | 10.60%             | 10.80%            | 17.90%            | 20.80%           |

Figure 27

|                           | S = 1000, D = 1000 | S = 500, D = 1000 | S = 1000, D = 500 | S = 500, D = 500 |
|---------------------------|--------------------|-------------------|-------------------|------------------|
| $p = 1\%$ , With Phase    | 0.00%              | 0.10%             | 3.00%             | 9.70%            |
| $p = 1\%$ , Without Phase | 74.20%             | 73.70%            | 74.60%            | 76.00%           |
| $p = 2\%$ , With Phase    | 0.00%              | 0.00%             | 0.00%             | 0.00%            |
| $p = 2\%$ , Without Phase | 13.90%             | 14.80%            | 41.00%            | 42.70%           |
| $p = 3\%$ , With Phase    | 0.00%              | 0.00%             | 0.00%             | 0.00%            |
| $p = 3\%$ , Without Phase | 0.00%              | 0.00%             | 5.70%             | 5.00%            |
| $p = 4\%$ , With Phase    | 0.00%              | 0.00%             | 0.00%             | 0.00%            |
| $p = 4\%$ , Without Phase | 0.00%              | 0.00%             | 0.00%             | 0.00%            |
| $p = 5\%$ , With Phase    | 0.00%              | 0.00%             | 0.00%             | 0.00%            |
| $p = 5\%$ , Without Phase | 0.00%              | 0.00%             | 0.00%             | 0.00%            |

Figure 26

|               | S = 1000, D = 1000 | S = 500, D = 1000 | S = 1000, D = 500 | S = 500, D = 500 |
|---------------|--------------------|-------------------|-------------------|------------------|
| With Phase    | 1.50%              | 8.80%             | 10.20%            | 22.30%           |
| Without Phase | 10.60%             | 10.80%            | 17.90%            | 20.80%           |

Figure 27

|                           | S = 1000, D = 1000 | S = 500, D = 1000 | S = 1000, D = 500 | S = 500, D = 500 |
|---------------------------|--------------------|-------------------|-------------------|------------------|
| $p = 1\%$ , With Phase    | 0.00%              | 0.10%             | 3.00%             | 9.70%            |
| $p = 1\%$ , Without Phase | 74.20%             | 73.70%            | 74.60%            | 76.00%           |
| $p = 2\%$ , With Phase    | 0.00%              | 0.00%             | 0.00%             | 0.00%            |
| $p = 2\%$ , Without Phase | 13.90%             | 14.80%            | 41.00%            | 42.70%           |
| $p = 3\%$ , With Phase    | 0.00%              | 0.00%             | 0.00%             | 0.00%            |
| $p = 3\%$ , Without Phase | 0.00%              | 0.00%             | 5.70%             | 5.00%            |
| $p = 4\%$ , With Phase    | 0.00%              | 0.00%             | 0.00%             | 0.00%            |
| $p = 4\%$ , Without Phase | 0.00%              | 0.00%             | 0.00%             | 0.00%            |
| $p = 5\%$ , With Phase    | 0.00%              | 0.00%             | 0.00%             | 0.00%            |
| $p = 5\%$ , Without Phase | 0.00%              | 0.00%             | 0.00%             | 0.00%            |

Figure 28A-28C

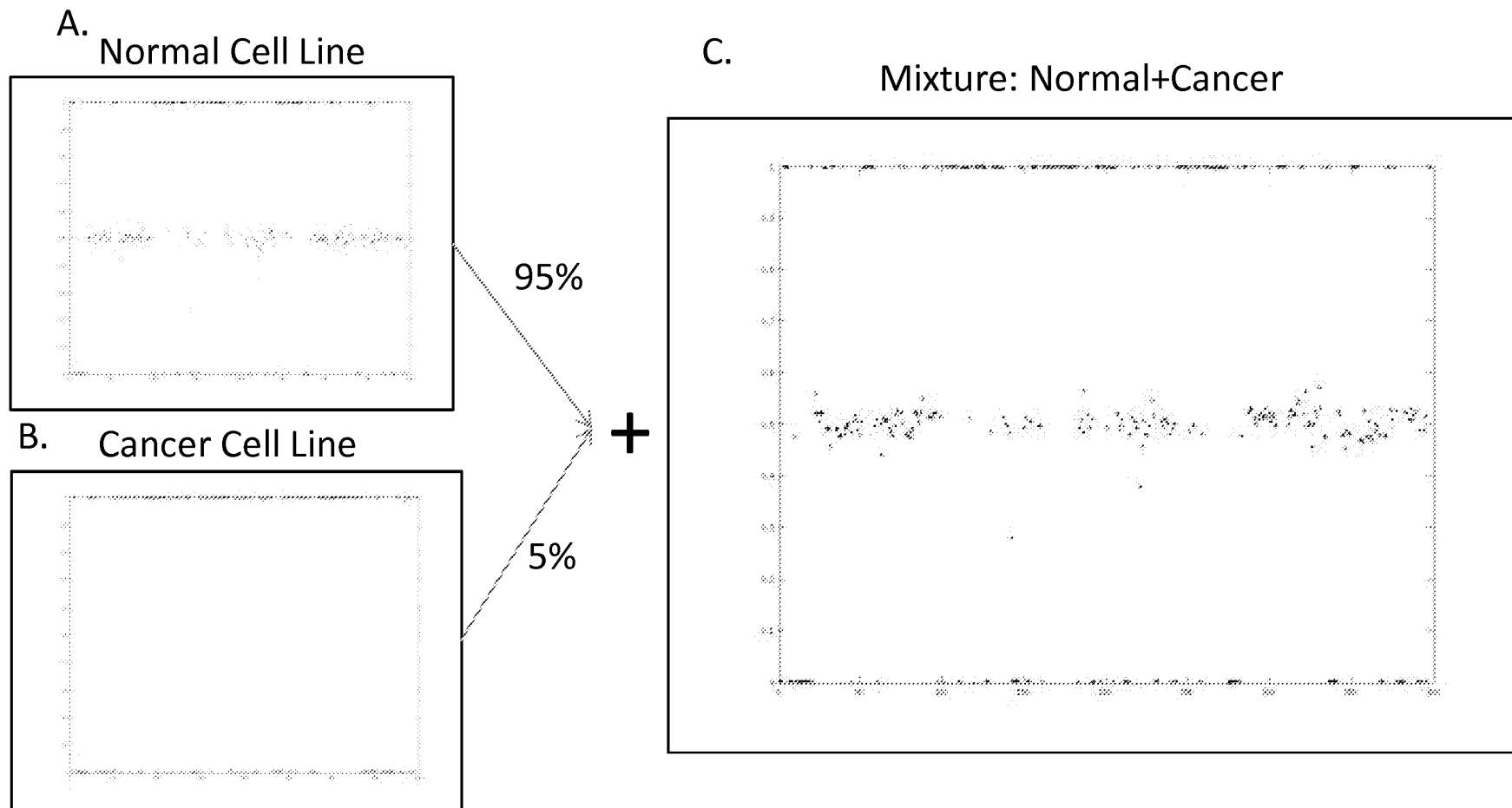


Figure 28A-28C

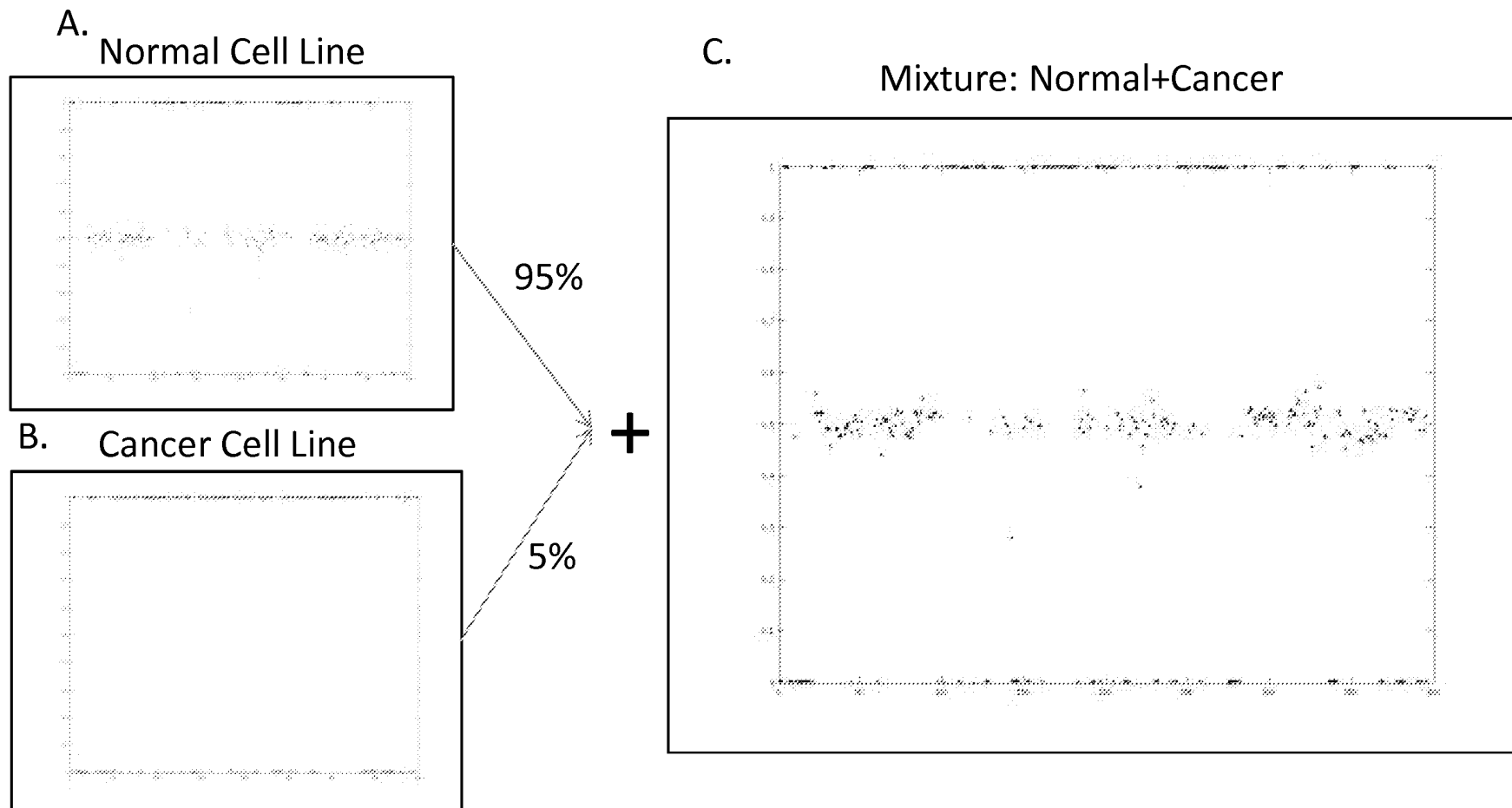


Figure 29

- Large tumor fraction (4.33%)

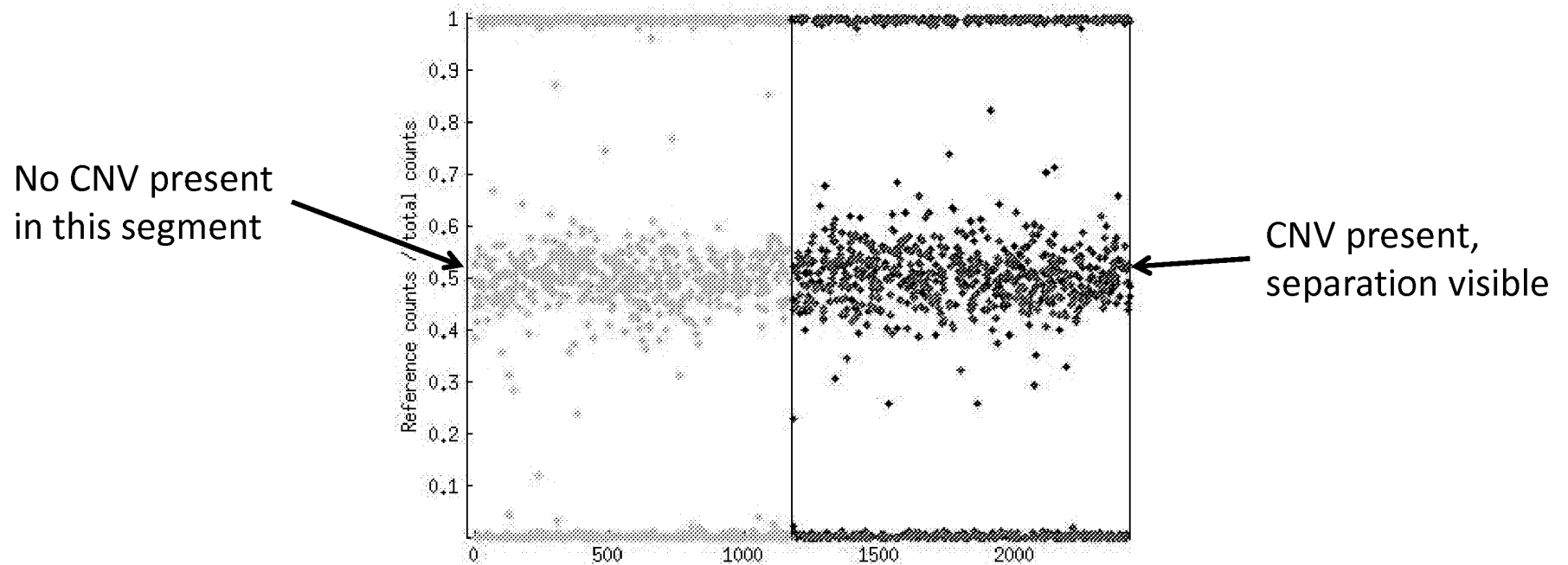


Figure 29

- Large tumor fraction (4.33%)

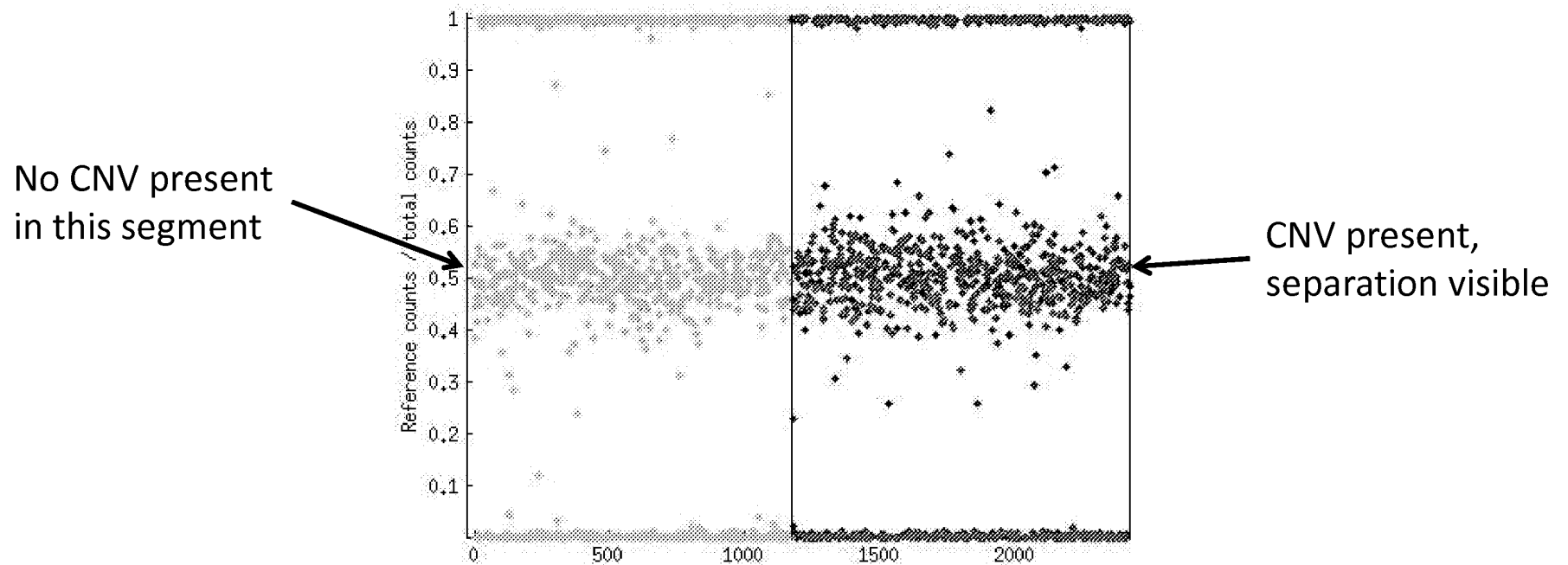


Figure 30

- Small tumor fraction (0.58%)

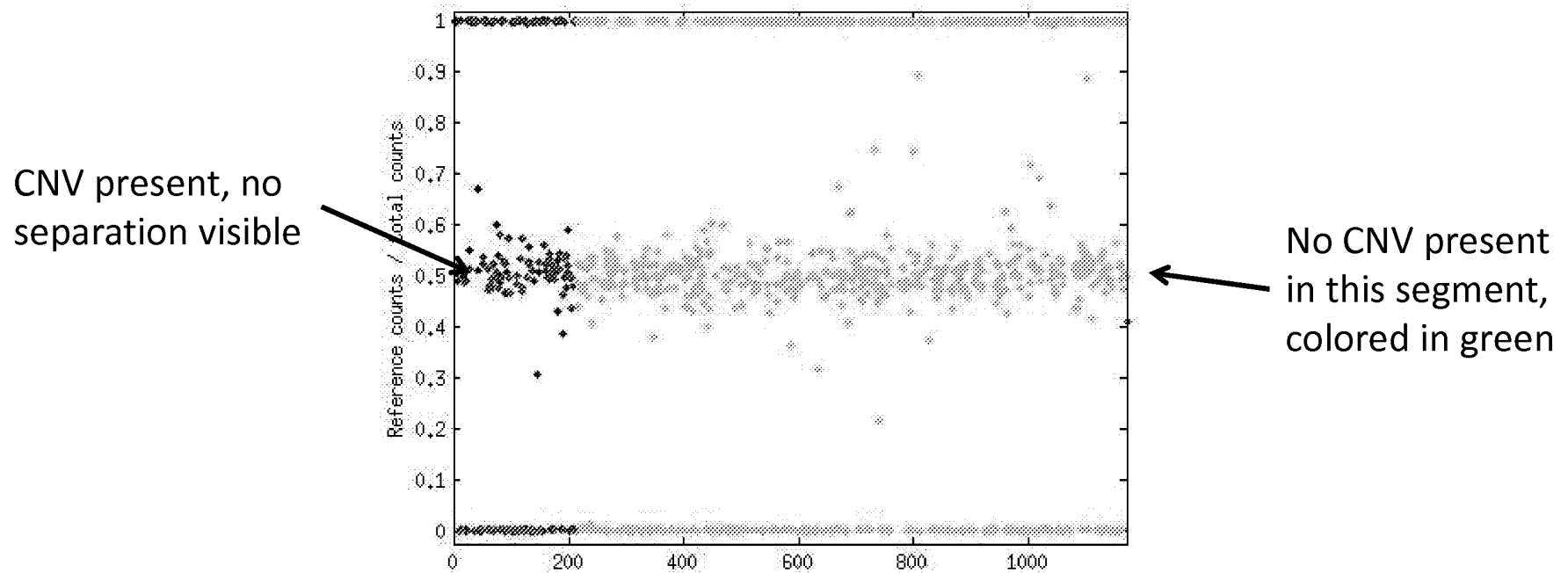
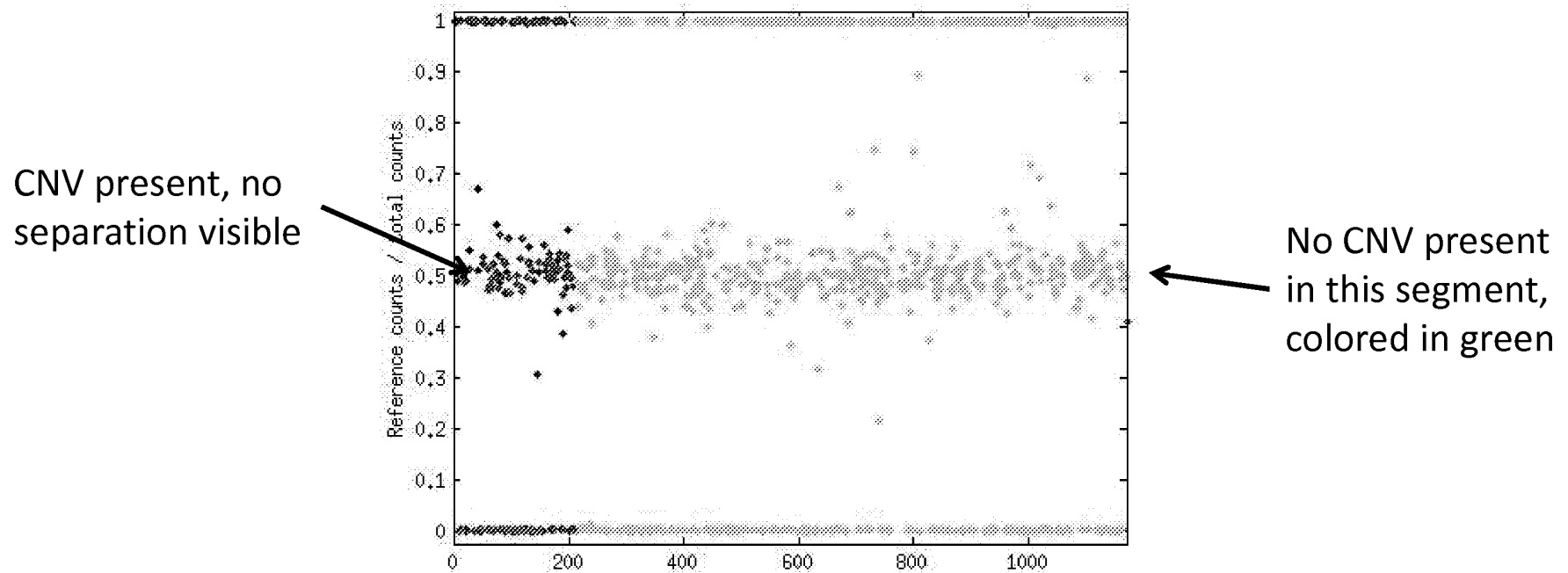


Figure 30

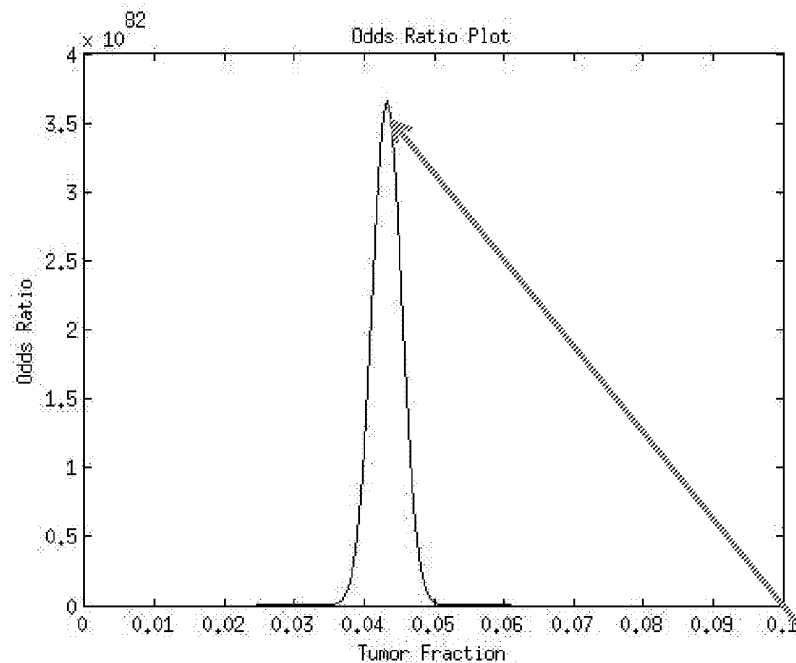
- Small tumor fraction (0.58%)



## Figures 31A and 31B

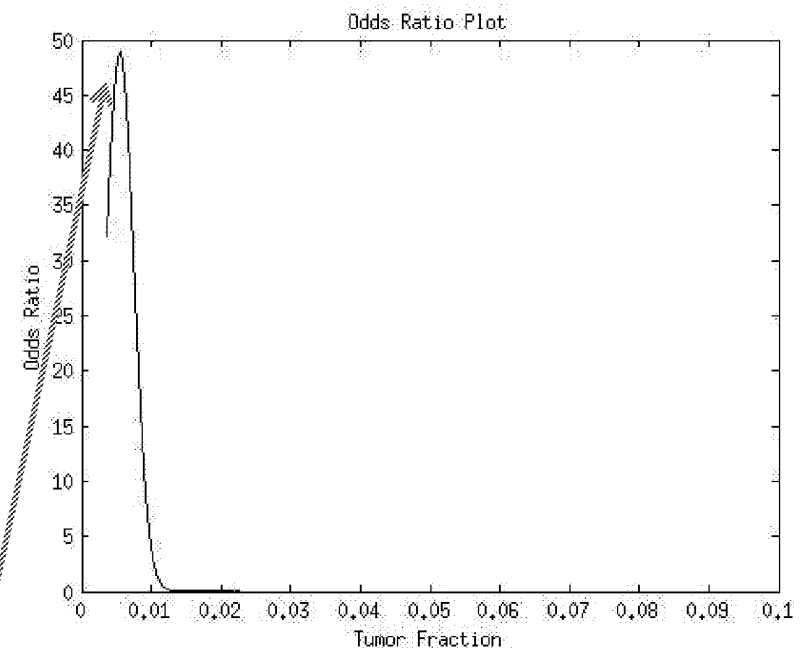
A.

Algorithm results for large tumor fraction



B.

Algorithm results for small tumor fraction

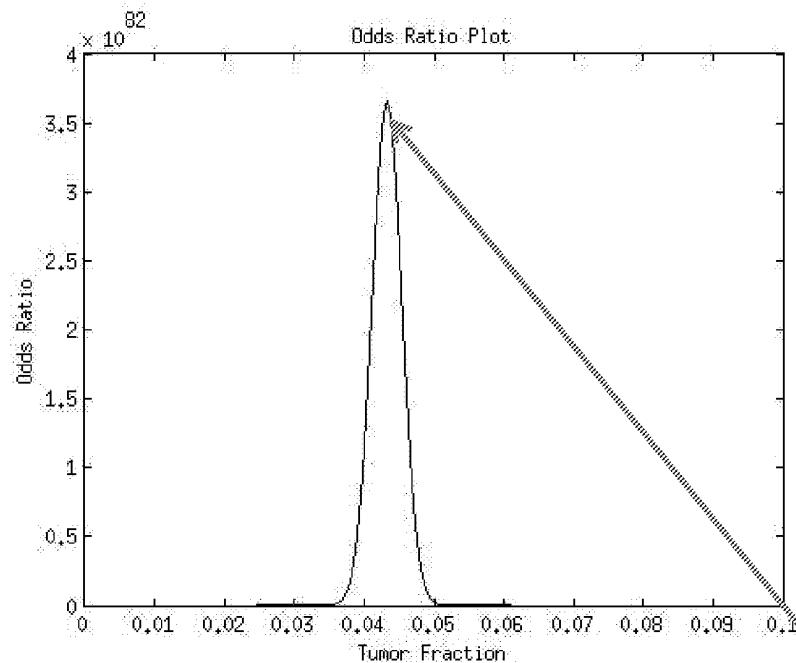


Maximum likelihood estimate of tumor fraction

# Figures 31A and 31B

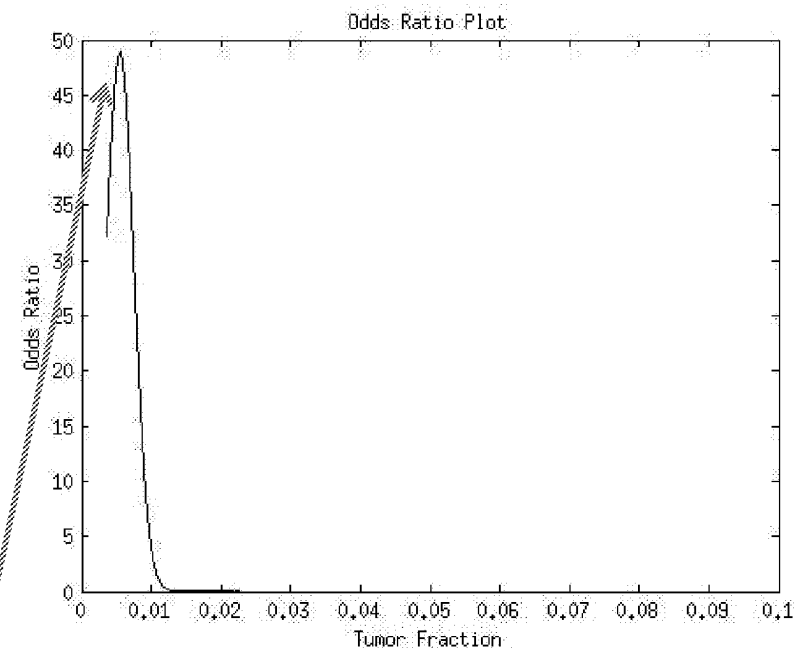
A.

Algorithm results for large tumor fraction



B.

Algorithm results for small tumor fraction

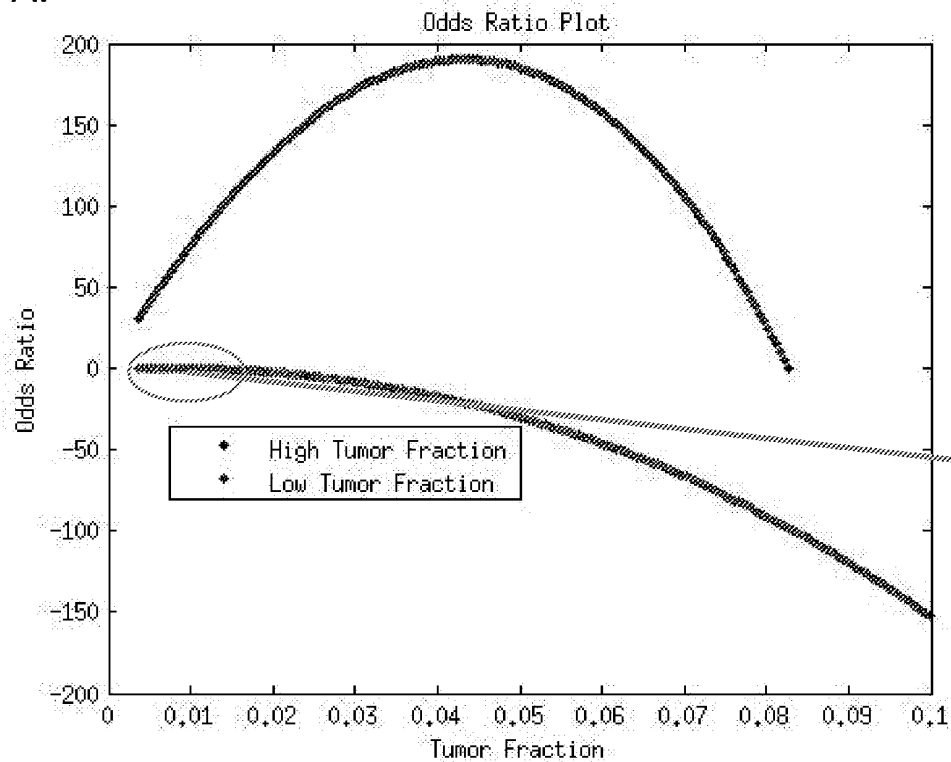


Maximum likelihood estimate of tumor fraction

## Figures 32A and 32B

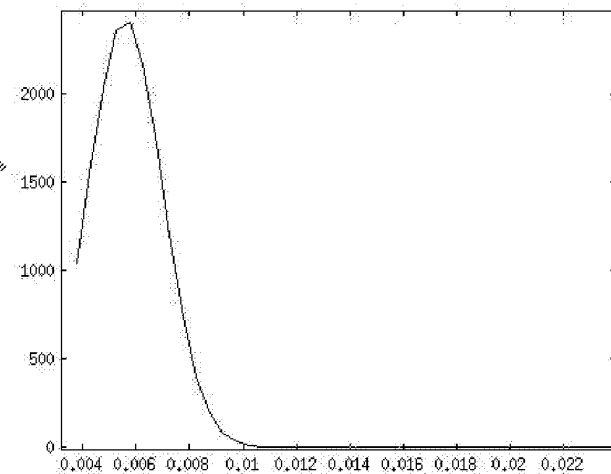
Compare the algorithm results for a large tumor fraction (4.33%) and a small tumor fraction (0.58%)

A.



B.

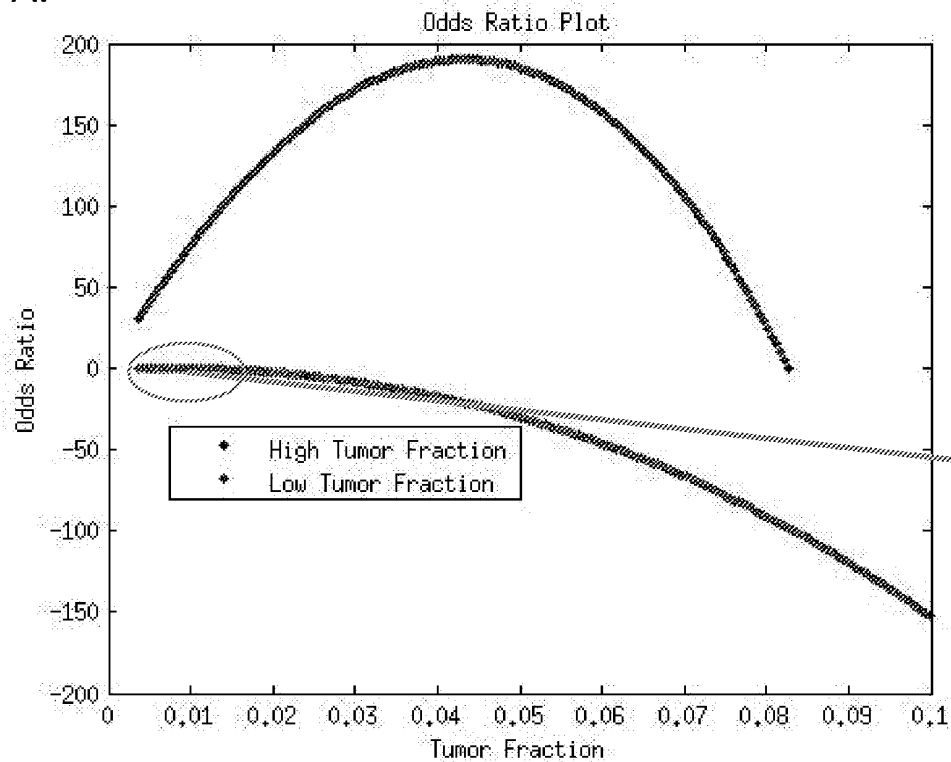
Small tumor results plotted  
In probability space



## Figures 32A and 32B

Compare the algorithm results for a large tumor fraction (4.33%) and a small tumor fraction (0.58%)

A.



B.

Small tumor results plotted  
In probability space

