(19) DANMARK

(10) **DK/EP 3294906 T3**

(12) Oversættelse af
europæisk patentskrift

Patent- og
Varemærkestyrelsen

(51) Int.Cl.: ***C 12 Q 1/6869 (2018.01)*** ***G 16 B 20/10 (2019.01)***

(45) Oversættelsen bekendtgjort den: **2024-08-05**

(80) Dato for Den Europæiske Patentmyndigheds
bekendtgørelse om meddelelse af patentet: **2024-07-10**

(86) Europæisk ansøgning nr.: **16724570.3**

(86) Europæisk indleveringsdag: **2016-05-10**

(87) Den europæiske ansøgnings publiceringsdag: **2018-03-21**

(86) International ansøgning nr.: **US2016031686**

(87) Internationalt publikationsnr.: **WO2016183106**

(30) Prioritet: **2015-05-11 US 201562159958 P**

(84) Designerede stater: **AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR**

(73) Patenthaver: **Natera, Inc., 201 Industrial Road , Suite 410, San Carlos, CA 94070, USA**

(72) Opfinder: **KIRKIZLAR, Huseyin Eser, 201 Industrial Road, Suite 410, San Carlos, California 94070, USA**
SALARI, Raheleh, 201 Industrial Road, Suite 410, San Carlos, California 94070, USA
SIGURJONSSON, Styrmir, 201 Industrial Road, Suite 410, San Carlos, California 94070, USA
ZIMMERMANN, Bernhard, 201 Industrial Road, Suite 410, San Carlos, California 94070, USA
RYAN, Allison, 201 Industrial Road, Suite 410, San Carlos, California 94070, USA
VANKAYALAPATI, Naresh, 201 Industrial Road, Suite 410, San Carlos, California 94070, USA

(74) Fuldmægtig i Danmark: **Potter Clarkson A/S, Regnbuepladsen 7, 1550 København V, Danmark**

(54) Benævnelse: **METHODS FOR DETERMINING PLOIDY**

(56) Fremdragne publikationer:
WO-A1-2012/108920
WO-A1-2013/130848
WO-A1-2014/018080
WO-A1-2015/164432
WO-A2-2012/142531
US-A1- 2011 301 854
US-A1- 2014 287 934
S. Y. SU ET AL: "Inferring combined CNV/SNP haplotypes from genotype data", BIOINFORMATICS, vol. 26, no. 11, 1 June 2010 (2010-06-01), pages 1437 - 1445, XP055199108, ISSN: 1367-4803, DOI: 10.1093/bioinformatics/btq157
DANIEL TALIUN ET AL: "Efficient haplotype block recognition of very long and dense genetic sequences", BMC BIOINFORMATICS, BIOMED CENTRAL, LONDON, GB, vol. 15, no. 1, 14 January 2014 (2014-01-14), pages 10, XP021171925, ISSN: 1471-2105, DOI: 10.1186/1471-2105-15-10
JIANNIS RAGOUSSIS: "Genotyping Technologies for Genetic Research", ANNUAL REVIEW OF GENOMICS

Fortsættes ...

AND HUMAN GENETICS, vol. 10, no. 1, 1 September 2009 (2009-09-01), pages 117 - 133, XP055116403, ISSN: 1527-8204, DOI: 10.1146/annurev-genom-082908-150116

DESCRIPTION

Description

FIELD OF THE INVENTION

[0001] The disclosed invention relates generally to methods of genetic analysis for determining chromosomal ploidy.

BACKGROUND OF THE INVENTION

[0002] 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 regions or entire chromosomes are associated with a variety of conditions, such as susceptibility or resistance to disease.

[0003] 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.

[0004] 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 can be more important in neurocognitive diseases than other forms of inherited mutations, including single nucleotide substitutions.

[0005] 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.

[0006] 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.

[0007] 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 CD16 cell surface immunoglobulin receptor) can increase susceptibility to systemic lupus erythematosus and similar inflammatory autoimmune disorders. WO-A-2015/164432 shows a method for determining ploidy of a chromosomal segment in which allelic frequency data from a set of polymorphic loci are received, phased allelic information is generated in order to establish the phase, individual probabilities of allelic frequencies for different ploidy states are generated, joint probabilities based on the Individual probabilities and the phased allelic information are generated, and selecting based on the joint probabilities a best fit model indicative of chromosomal ploidy.

[0008] Thus, improved methods are needed to detect deletions and duplications of chromosome regions 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

[0009] The invention provides a method for determining ploidy of a chromosomal segment in a sample of an individual as set out in claims 1-14. Specifically, the invention is a method for determining ploidy of a chromosomal segment in a sample of an individual, the method comprising: a) receiving allele frequency data for each SNP of a set of SNPs comprising 200 SNPs on a plurality of subsegments within the chromosomal segment, wherein within each subsegment 95% of pairwise SNP comparisons between any two SNPs within that chromosome/region have |D' | of > 95% , wherein the allele frequency data comprises the amount of each allele present in the sample at each SNP; b) generating phased allelic information for the set of SNPs by estimating the phase of the genotypic measurement data, taking into account an increased statistical correlation of SNPs within the same subsegment; c) generating individual likelihoods of allele frequencies for the set of SNPs for different ploidy states using the allele frequency data; d) generating joint likelihoods for the set of linked SNPs using the individual likelihoods and the phased allelic information; and e) selecting, based on the joint likelihoods, a best fit model indicative of chromosomal ploidy, thereby determining the ploidy of the chromosomal segment.

[0010] Disclosed are methods, for detecting cancer or a chromosomal abnormality in a gestating fetus. The methods can utilize a set of SNPs that are found within haploblocks and can include analyzing a series of target chromosomal regions related to CNV in cancer or a chromosomal abnormality in a gestating fetus.

[0011] Disclosed is a method for determining ploidy (i.e. copy number) of a chromosome or chromosomal region of interest (i.e. target chromosomal region) in a sample from an individual. The method can include the following steps:

1. a. making genotypic measurements for 200 on the chromosome or chromosome region of interest from a sample of blood, or a fraction thereof from the target individual, wherein at least 95% of the polymorphic loci of the plurality of polymorphic loci (or SNPs from the set of SNPs) have strong linkage disequilibrium with at least one other polymorphic loci of the plurality of polymorphic loci or SNP of the set of SNPs;
2. b. estimating the phase of the genotypic measurements; and

wherein the method employs the features defined in the claims.

[0012] The step of making genotypic measurements can be done by measuring genetic material using techniques selected from the group consisting of padlock probes, circularizing probes, genotyping microarrays, SNP genotyping assays, chip based microarrays, bead based microarrays, other SNP microarrays, other genotyping methods, Sanger DNA sequencing, pyrosequencing, high throughput sequencing, reversible dye terminator sequencing, sequencing by ligation, sequencing by hybridization, other methods of DNA sequencing, other high throughput genotyping platforms, fluorescent in situ hybridization (FISH), comparative genomic hybridization (CGH), array CGH, and multiples or combinations thereof. Genotypic measurements can be performed using high throughput sequencing or genotyping microarrays.

[0013] A method for determining ploidy of a chromosome or chromosomal region of interest (i.e. target chromosomal region) in a sample of an individual may include the following steps:

1. a. receiving allele frequency data for each SNP of a set of SNPs that includes 200SNPs on a plurality of segments within the chromosomal region, wherein each segment

comprises loci with strong linkage disequilibrium (e.g. haploblocks), wherein the allele frequency data comprises the amount of each allele present in the sample at each loci;

2. b. generating phased allelic information for the set of SNPs by estimating the phase of the allele frequency data taking into account an increased statistical correlation of polymorphic loci within the same segment;
3. c. generating individual likelihoods of allele frequencies for the polymorphic loci for different ploidy states using the allele frequency data;
4. d. generating joint likelihoods for the plurality of linked polymorphic loci using the individual likelihoods and the phased allele frequency data; and
5. e. selecting, based on the joint likelihoods, a best fit model indicative of chromosomal copy number, thereby determining the copy number of the chromosome or chromosome region.

[0014] In the method for determining ploidy, set out above, at least 95% of the polymorphic loci of the plurality of SNPs from the set of SNPs have strong linkage disequilibrium with at least one other SNP of the plurality of loci (e.g. set of SNPs). The method can detect CNV for example, by detecting an AAI above a sensitivity or cutoff value.

[0015] In the method for determining ploidy set out above, receiving allele frequency data can include receiving nucleic acid sequencing data for 200 amplicons spanning each loci of the plurality of polymorphic loci and generating the allele frequency data from the sequencing data.

[0016] The method for determining ploidy set out above, can further include the following;

1. a. amplifying the plurality of polymorphic loci (e.g. set of SNPs) by an amplification method that includes the following:
   1. i. forming a reaction mixture that includes circulating free nucleic acids derived from the sample, a polymerase and a pool of primers comprising at least 200, 250, 300, 400, 500, 1,000, 2,500, 5,000, or 10,000 primers or primer pairs that each specifically bind to a primer binding sequence located within an effective distance of one of the polymorphic loci; and
   2. ii. subjecting the reaction mixture to amplification conditions, thereby generating a plurality of amplicons; and

subjecting each of the amplicons to a nucleic acid sequencing reaction to generate the nucleic acid sequencing data for the amplicons.

[0017] In addition to the above, methods of amplifying, reaction mixtures, and compositions comprising a set, pool, or plurality of primers or primer pairs are provided herein, that includes at least 200, 250, 500, 1000, or 2,500 primers or primer pairs, or between 100, 200, 250, 500, 1000, 2,500, 5,000, or 10,000 on the low end of the range, and 250, 500, 1000, 2,500, 5,000, or 10,000 on the high end of the range, that each specifically bind to a primer binding sequence located within one or more of a plurality of haploblocks, wherein each haploblock comprises at least 2, 3, 4, 5 or 10 of the primer binding sequences and wherein at least 50, 75, 90, 95, or 100% of the primer binding sequences are located within haploblocks.

[0018] A reaction mixture provided herein, can include:

1. a. a population of circulating free nucleic acids from an individual, or nucleic acid fragments derived therefrom, and
2. b. a composition that includes at least 200, 250, 500, 1000, 2,500, 5,000, or 10,000 primers or primer pairs that each specifically bind to a primer binding sequence located within one or more of a plurality of haploblocks, wherein each haploblock comprises at least 2, 3, 4, 5 or 10 of the primer binding sequences and wherein at least 50, 75, 90, 95, or 100% of the primer binding sequences are located within haploblocks.

[0019] The primer binding sequences can be found within a chromosome region known to exhibit copy number variation (CNV) associated with a disorder or disease, such as cancer.

[0020] Further embodiments and aspects of the invention are provided in the detailed description section.

Definitions

[0021] *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 regions, such as the presence of deletions or duplications of a chromosome region) 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. The deletion of a chromosome region may be a microdeletion.

[0022] *Allelic Data* refers to a set of genotypic data for 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.

[0023] *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 thus providing allele frequency data. If each allele is counted in a binary fashion, then the allele count will be a whole number. If the alleles are counted probabilistically, then the allele count can be a fractional number.

[0024] *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 can be treated probabilistically, that is, the likelihood that a given allele is present for a give sequence read is a fraction between 0 and 1, or they can be treated in a binary fashion, that is, any given read is considered to be exactly zero or one copies of a particular allele.

[0025] *Allelic imbalance* for aneuploidy determinations, such as CNV determinations, refers to the difference between the frequencies of the alleles for a locus. It is an estimate of the difference in the copy of numbers of the homologs. Allelic imbalance can arise from the complete loss of an allele or from an increase in copy number of one allele relative to the other. Allelic imbalances can be detected by measuring the proportion of one allele relative to the other in cells from individuals that are constitutionally heterozygous at a given locus. (Mei et al, Genome Res, 2000). The proportion of abnormal DNA for a CNV can be measured by the average allelic imbalance (AAI), defined as $|(H1 - H2)|/(H1 + H2)$, where Hi is the average number of copies of homolog i in the sample and $Hi/(H1 + H2)$ is the fractional abundance, or homolog ratio, of homolog i. The maximum homolog ratio is the homolog ratio of the more abundant homolog.

[0026] *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.

[0027] *Haplotype Data*, also "Phased Data" or "Ordered Genetic Data," refers to data from a single chromosome or chromosome region in a diploid or polyploid genome, e.g., either the segregated maternal or paternal copy of a chromosome in a diploid genome.

[0028] *Linkage Disequilibrium* (LD) refers to the non-random association of alleles at two loci that can be measured by r2, |D|, |D'|. Two loci with high disequilibrium are said to be in "strong LD", or have "weak recombination". Haplotype blocks are sets of consecutive sites between which there is little or no evidence of historical recombination. Based on the default haplotype block definition (Gabriel et al, Science, 2002) a block is created by identifying a set of SNP loci on the same chromosome for which 95% of pairwise SNP comparisons between any 2 SNPs within that chromosome region have |D'| > 95%. Therefore, for the purposes of the present disclosure, polymorphic loci (e.g. SNPs) are said to have strong linkage disequilibrium if 95% of pairwise SNP comparisons between any two SNPs within that chromosome/region have |D'| of > 95%.

[0029] *Phasing* refers to the act of estimating the haplotypic genetic data of an individual. It may refer to the act of estimating which of the two alleles at a locus are associated with each of the two homologous chromosomes in an individual. "Perfect haplotyping" in discussions herein in the context of methods for analyzing a sample that includes ctDNA, is used to refer to molecular haplotyping through a supplementary tumor sample. Methods provided herein are especially well-suited for imperfectly phased data, especially data whose haplotype or phase has been estimated using an algorithm.

[0030] *Phased Data* refers to genetic data where one or more haplotypes have been estimated.

[0031] *Copy Number Hypothesis*, also "Ploidy State Hypothesis," refers to a hypothesis concerning the number of copies of a chromosome or chromosome region 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 regions, if any, from a related individual correspond genetically to a given chromosome from an individual.

[0032] *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.

[0033] *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.

[0034] *Haploblocks* or *haplotype blocks* refers to a segment of a chromosome that contains a set of consecutive loci between which there is little or no evidence of historical recombination. Based on the default haplotype block definition (Gabriel et al, Science, 2002) a block is created if 95% of pairwise SNP comparisons are "strong LD" using a 95% r2 cutoff. Publically available programs, such as plink (v1.90b3p 64-bit (10 Oct 2014)), can be used to identify known or identified haploblocks for regions of interest based on this definition. It is noteworthy for considerations herein, that a series of consecutive SNPs that are amplified or deleted together are considered to be in the same haploblock, when haplotyping is done using a tumor sample.

[0035] Where ranges of values have been given in this disclosure, all intermediate values and end-points of the range form part of the disclosure.

[0036] Other features and advantages of the invention will be apparent from the following detailed description and from the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0037] The present disclosure 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 methods.

FIG. 1 is an example of CNV region identification. Illustrated is chromosome 8. The x-axis represents the genomic position on the chromosome. Both the x-axis range (0 - 250 Mb) and the y-axis range (0 - 453 patients) is consistent across plots.

FIG. 2 is a bar chart of 14 prioritized candidate regions with chromosome number and positions labeled on the x-axis and cumulative patient coverage on the y-axis.

FIGS. 3A-3B is a table of correlations of CNV events for 14 patients.

FIGS. 4A-4H are graphs of exemplary CNV region identification: FIG. 4A PIK3CA, chromosome 3; FIG. 4B MYC, chromosome 8; FIG. 4C KRAS, chromosome 12; FIG. 4D RB1, chromosome 13, FIG. 4E CDH1, chromosome 16, FIG. 4F MAP2K4 and NF1, chromosome 17; FIG. 4G AKT2, chromosome 19; and FIG. 4H, chromosome 20.

FIG. 5 shows an example system architecture X00 useful for performing embodiments of the present invention.

FIG. 6 illustrates an example computer system for performing embodiments of the present invention.

FIG. 7 is a graph of haploblock size for target chromosome regions of 8 target lung cancer-associated genes analyzed in Example 5.

FIG. 8 provides a table of AAI as a function of TCF and tumor copy number, and detection limit of different technologies

DETAILED DESCRIPTION OF THE INVENTION

[0038] Disclosed are improved methods for determining ploidy of a chromosomal segment in a sample of an individual as set out in claims 1-14. Specifically, the invention is a method for determining ploidy of a chromosomal segment in a sample of an individual, the method comprising: a) receiving allele frequency data for each SNP of a set of SNPs comprising 200 SNPs on a plurality of subsegments within the chromosomal segment, wherein within each subsegment 95% of pairwise SNP comparisons between any two SNPs within that chromosome/region have ID' | of > 95% , wherein the allele frequency data comprises the amount of each allele present in the sample at each SNP; b) generating phased allelic information for the set of SNPs by estimating the phase of the genotypic measurement data, taking into account an increased statistical correlation of SNPs within the same subsegment; c) generating individual likelihoods of allele frequencies for the set of SNPs for different ploidy states using the allele frequency data; d) generating joint likelihoods for the set of linked SNPs using the individual likelihoods and the phased allelic information; and e) selecting, based on the joint likelihoods, a best fit model indicative of chromosomal ploidy, thereby determining the ploidy of the chromosomal segment.. In methods for detecting the presence or absence of CNV, haplotype information is estimated using analytical methods. By choosing polymorphic loci, and designing primers and assays for amplifying the same, that are within haplotype blocks, or haploblocks, informatics haplotyping can be improved. This can be especially beneficial when used as part of CNV detection methods, where the haploblocks are within chromosome regions known to exhibit CNV correlated with disease, such as cancer. Accordingly, for cfDNA samples by choosing polymorphic loci, and designing primers and assays for amplifying the same, additional sampling, such as from a buccal sample or from a tumor sample becomes unnecessary. It is noted that chromosomal regions that are deleted or duplicated in some diseases or disorders, such as cancer, can be referred to as chromosome segments. These chromosomal segments are typically made up of numerous segments of loci that share high linkage disequilibrium with neighboring loci.

[0039] Disclosed are methods of determining the presence or absence of copy number variations, such as deletions or duplications of chromosome regions. The methods are particularly useful for detecting small deletions or duplications, which can be difficult to detect with high specificity and sensitivity using prior art methods due to the small amount of data available from the relevant chromosome region. The methods include improved analytical methods, improved bioassay methods, and combinations of improved analytical and bioassay methods. The methods may be adapted 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 in circulating DNA fractions and/or 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 as well as with developmental disorders.

[0040] 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. Furthermore, biopsies are limited to detecting mutations in the biopsy section sampled, not the entire tumor. 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 can 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 can be used increase throughput and reduces costs.

[0041] The disclosed methods can be 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 can be analyzed. The sample can be from a tissue or organ suspected of having a deletion or duplication such as cells or a mass suspected of being cancerous. The methods can be used to detect deletion(s) or duplication(s) that are only present in one cell or a small number of cells in a mixture containing cells with the deletion(s) or duplication(s) and cells without the deletion(s) or duplication(s). In illustrative embodiments, cfDNA or cfRNA from a blood sample, or a fraction thereof, from the individual is analyzed according to methods provided herein. cfDNA or cfRNA can be secreted by cells, for example, cfDNA or cfRNA can be released by cells undergoing necrosis or apoptosis, such as cancer cells. The methods can be used to detect deletions or duplications that are only present in a small percentage of the cfDNA or cfRNA.

[0042] The methods can be used for non-invasive or invasive prenatal testing of a fetus by determining the presence or absence of deletions or duplications of a chromosome region or an entire chromosome, such as deletions or duplications known to be associated with severe mental or physical handicaps, learning disabilities, or cancer. For non-invasive prenatal testing (NIPT), cells, cfDNA or cfRNA from a blood sample, or a fraction thereof, from the pregnant mother can be 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. The Examples section herein, provides exemplary methods that focus on detecting CNV in cancer. However, a skilled artisan will understand that these methods as they relate to CNV in cancer, can be used for determining chromosomal ploidy in NIPT, where only imperfect haplotyping is performed, especially when haplotyping is not performed on a tumor sample. Chromosomes and chromosome regions that are duplicated or deleted in NIPT are known, and methods disclosed herein can be used to determine haploblocks, and design primers, primer pairs, and assays for determining alleles within polymorphic loci in those haploblocks.

[0043] 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 a 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.

Methods for Determining Ploidy

[0044] The disclosed methods are based in part on the finding that the ability to detect aneuploidy of a chromosome region(s) can be improved by selecting polymorphic loci within segments, called haploblocks or haplotype blocks, within the chromosome regions where neighboring SNPs demonstrate strong linkage disequilibrium. The improvements in AAI, copy number or ploidy determination and aneuploidy or CNV detection are especially pronounced when a pool of primers are selected for determining the allele frequency at a set of SNPs within a plurality of haploblocks within a target chromosome region, and the method includes a step where the phase of the allele frequency data within a chromosome region of interest is estimated to generate imperfect haplotype data that is used for the ploidy determination or aneuploidy detection.

[0045] A method for determining ploidy (i.e. copy number) of a chromosome or chromosomal region of interest (i.e. target chromosomal region) in a sample from an individual can include the following steps:

1. a. making genotypic measurements for 200 SNPs on the chromosome or chromosome region of interest from a sample of blood, or a fraction thereof from the target individual, wherein at least 95% of the SNPs from the set of SNPs have strong linkage disequilibrium with at least one other SNP of the set of SNPs;
2. b. estimating the phase of the genotypic measurements; and
3. c. determining on a computer, the likelihood of different ploidy states of the chromosome or chromosome region of interest by comparing the phased genotypic measurements to a set of joint distribution models of expected genotypic measurements for different ploidy states using identified chromosome crossover locations, thereby determining the ploidy state as the copy number of the chromosome or chromosome region with the highest likelihood.

[0046] The determining can be performed by:

1. a. creating, on a computer, a set of ploidy state hypothesis where each ploidy state hypothesis is one possible ploidy state of the [target] chromosome or chromosome region of interest;

2. b. building a set of joint distribution models of expected genotypic measurements at the set of SNPs for each hypothesis given identified chromosome crossover locations;
3. c. determining, on the computer, the likelihood of each of the hypotheses given the estimated phase of the genotypic measurements and the joint distribution model.

[0047] The step of making genotypic measurements can be done by measuring genetic material using techniques selected from the group consisting of padlock probes, circularizing probes, genotyping microarrays, SNP genotyping assays, chip based microarrays, bead based microarrays, other SNP microarrays, other genotyping methods, Sanger DNA sequencing, pyrosequencing, high throughput sequencing, reversible dye terminator sequencing, sequencing by ligation, sequencing by hybridization, other methods of DNA sequencing, other high throughput genotyping platforms, fluorescent in situ hybridization (FISH), comparative genomic hybridization (CGH), array CGH, and multiples or combinations thereof. Genotypic measurements can be performed using high throughput sequencing or genotyping microarrays.

[0048] , The step of measuring genetic material can be performed on genetic material that is amplified prior to being measured, using a technique that is selected from Polymerase Chain Reaction (PCR), ligand mediated PCR, degenerative oligonucleotide primer PCR, Multiple Displacement Amplification (MDA), allele-specific PCR, allele-specific amplification techniques, bridge amplification, padlock probes, circularizing probes, and combinations thereof. The amplification can be performed using multiplex PCR including PCR using the sets of primers the pools, sets, pluralities, or libraries of primers set out herein.

[0049] Also disclosed is a method for determining AAI, ploidy (i.e. copy number) or detecting copy number variation (CNV) or aneuploidy, of a chromosome or chromosomal region of interest (i.e. target chromosomal region) in a sample of an individual. The method includes the following steps:

1. a. measuring and/or receiving allele frequency data for each loci (e.g. SNP) of a plurality of polymorphic loci (e.g. set of SNPs) that includes at least 200, 250, 300, 400, 500, 1,000, 2,500, 5,000, or 10,000 loci (e.g. SNPs) on a plurality of segments within the chromosomal region, wherein each segment comprises loci with strong linkage disequilibrium (e.g. haploblocks), wherein the allele frequency data comprises the amount of each allele present in the sample at each loci;
2. b. estimating the phase of the allele frequency data taking into account an increased statistical correlation of polymorphic loci within the same segment;
3. c. generating individual likelihoods of allele frequencies for the polymorphic loci for different ploidy states using the allele frequency data;
4. d. generating joint likelihoods for the plurality of linked polymorphic loci using the individual likelihoods and the phased allele frequency data; and
5. e. selecting, based on the joint likelihoods, a best fit model indicative of chromosomal copy number, thereby determining the copy number of the chromosome or chromosome

region.

[0050] In the method for determining ploidy, set out above, at least 50, 60, 70, 75, 80, 90, 95, 96, 97, 97, 99, or 100% of the polymorphic loci of the plurality of polymorphic loci (or SNPs from the set of SNPs) can have strong linkage disequilibrium with at least one other loci (e.g. SNP) of the plurality of loci (e.g. set of SNPs).

[0051] In the method for determining ploidy set out above, receiving allele frequency data can include receiving nucleic acid sequencing data for at least 200, 250, 300, 400, 500, 1,000, 2,500, 5,000, or 10,000 amplicons spanning each loci of the plurality of polymorphic loci and generating the allele frequency data from the sequencing data.

[0052] The method for determining ploidy set out above, can further include the following;

1. a. amplifying the plurality of polymorphic loci (e.g. set of SNPs) by an amplification method that includes the following:
    1. i. forming a reaction mixture that includes circulating free nucleic acids derived from the sample, a polymerase and a pool of primers comprising at least 200, 250, 300, 400, 500, 1,000, 2,500, 5,000, or 10,000 primers or primer pairs that each specifically bind to a primer binding sequence located within an effective distance of one of the polymorphic loci; and
    2. ii. subjecting the reaction mixture to amplification conditions, thereby generating a plurality of amplicons; and
2. b. subjecting each of the amplicons to a nucleic acid sequencing reaction to generate the nucleic acid sequencing data for the amplicons.

[0053] For any of the quantitative, allelic methods provided herein, a confidence may be computed for the copy number determination. In combination with any 1 or more of the illustrative optional additional steps set out herein in this paragraph, or as a separate example, the method can further include obtaining prior likelihoods of each hypothesis from population data, and computing the confidence using Bayes Rule. In combination with any 1 or more of the illustrative embodiments set out herein in this paragraph, or as a separate example, the method can further include calculating a platform response to statistically correct for bias and/or increase the accuracy of the genotypic measurements, wherein the platform response is a mathematical characterization of the input/output characteristics of a genetic measurement platform. In combination with any 1 or more of the illustrative embodiments set out herein in this paragraph, or as a separate example, an average allelic imbalance can calculated and wherein the copy number determination is indicative of a copy number variation if the average allelic imbalance is equal to or greater than a cutoff value, which can be a sensitivity for an assay method, such as an AAI of 0.45%. In combination with any 1 or more of the illustrative embodiments set out herein in this paragraph, or as a separate example, a likelihood for each ploidy state can be determined based on a beta binomial distribution of expected and observed genetic or allelic frequency data at the plurality of SNP loci. In combination with any 1 or more of the illustrative embodiments set out herein in this paragraph, or as a separate example, the determining can include determining the ploidy state with the highest likelihood based on Bayesian estimation, as an indication of the number of copies of the chromosome or chromosome region of interest. In combination with any 1 or more of the illustrative embodiments set out herein in this paragraph, or as a separate example, the sample is the only sample whose phase is estimated for the individual. Thus, in these illustrative embodiments, for example wherein the subject is suspected of having cancer, the phase of genetic material in a tumor sample is not estimated. In these illustrative embodiments, for example wherein the subject is a pregnant mother, the phase of a genetic material from another sample from the mother besides a plasma sample, is not estimated.

[0054] Methods provided herein can include analysis of at least 200, 250, 300, 350, 400, 500, 600, 700, 750, 800, 900, 1000, 1250, 1500, 1750, 2000 polymorphic loci, such as SNPs , on the low end of the range, and 200, 250, 300, 350, 400, 500, 600, 700, 750, 800, 900, 1000, 1250, 1500, 1750, 2000, 2500, 5000, or 10,000 polymorphic loci, such as SNPs, on the high end of the range located within a plurality of haploblocks on a target chromosome region, each haploblock having between 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 25, or 50 SNP loci on the low end of the range, and 5, 6, 7, 8, 9, 10, 20, 25, 50 75, 100, 150, 200, or 250 SNP loci on the high end of the range. Furthermore, a plurality of haploblocks on the same chromosome region analyzed in such methods can include, for example, between 2, 3, 4, 5, 6, 7, 8, 9, or 10 haploblocks per target chromosome or chromosome region on the low end of the range, and 5, 6, 7, 8, 9, 10, 20, 25, 50 75, 100, 150, 200, or 250 haploblocks per chromosome or chromosome region on the high end of the range. A skilled artisan will understand that the size of the chromosome region will influence the number of haploblocks and SNPs within haploblocks, for that chromosome region. This is illustrated in the Examples herein, where the target chromosomal regions that were identified for lung cancer were small than those identified for ovarian cancer. Thus, less SNPs per chromosome region that occur within haploblocks of, for example at least 5 loci, were identified in the lung cancer regions than the ovarian cancer regions analyzed. This improvement in aneuploidy detection is especially valuable for samples in which only a small percentage of the nucleic acids in the sample exhibit aneuploidy, such as a plasma sample with circulating fetal or tumor DNA. The power of the analysis is further apparent when considering that aneuploidy occurs in fetal disorders and many cancers, such as ovarian or lung cancer, in targeted regions of the genome.

[0055] Accordingly, target segments are identified by identifying segments that include polymorphic loci with strong linkage disequilibrium using a 70, 75, 80, 85, 90, 95 or 99% |D'| cutoff where 95, or 99% of pairwise SNP comparisons show a strong linkage disequilibrium. The segments may be haploblocks (i.e. 95% of pairwise SNP comparisons are "strong LD" using a |D'| > 95% cutoff). SNPs with minor allele frequency of less than 5.0, 10.0, 15.0 and 20.0% can be ignored in illustrative examples of the method. Haploblocks can include blocks of at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 75, 80, 90, 100, 200, 250, 500, or 1000 neighboring SNPs that show a strong linkage disequilibrium.

[0056] Programs are known in the art for estimating haploblocks. For example, the program called plink (available on the Internet at pngu.mgh.harvard.edu) can be used to estimate haploblocks, as illustrated in the Examples section herein. The program estimates haploblocks for a given set of SNPs based on a given reference panel. Other programs are publicly available for estimating haploblocks, in addition to plink include LDHat (availableon the Internet at ldhat.sourceforge.net/), Haploview (available on the Internet at www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview), LdCompare (available on the Internet at www.affymetrix.com/support/developer/tools/devnettools.affx), TASSEL (available on the Internet at www.maizegenetics.net/?Itemid=119&id=89&option=com_content&task=view), and rAggr (available on the Internet at raggr.usc.edu).

[0057] This disclosure provides guidelines for assay design parameters for detecting polymorphic loci. For example, proper assay designs in illustrative examples are based on selecting non-interactive assays within chromosomal regions that show a high percentage of aneuploidy covering at least 50% of the chromosomal region. Furthermore, for cancer detection, recurrence profiles can be analyzed, such as that shown in FIG. 1. Finally, chromosomal regions that exhibit a minor allele frequency of 10-50% can be chosen. Guidelines for identifying amplicons and amplification parameters are provided herein. For example,
amplicons can be identified that include SNPs, with lengths between 50 and 75 bp, with a Tm of between 53-59C with a GC content of 30-70 and with MAF of 10-50%.

[0058] Data generated by a method of the invention, for determining ploidy. that takes into account an increased probability of linkage for loci found in haploblocks located within a chromosome region, can be combined with any analytical method that uses imperfectly haplotyped allele data at polymorphic loci to determine ploidy to improve the accuracy and sensitivity of such ploidy analysis.

[0059] 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. These predictions can be further improved by taking advantage of the increased statistical association for SNPs within haploblocks, by choosing SNP loci to analyze, that are within haploblocks.

Samples

[0060] 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 can 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.

[0061] 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 can 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 can be a mixture of DNA derived from target cells (such as cancer cells) and non-target cells (such as non-cancer cells).

[0062] The sample can contain or can be suspected of containing a mixture of DNA (or RNA), such as mixture of cancer DNA (or RNA) and noncancerous DNA (or RNA). At least 0.5, 1, 3, 5, 7, 10, 15, 20, or 25% of the cells in the sample can be cancer cells. In other examples, at least 0.5, 1, 3, 5, 7, 10, 15, 20, or 25% of the DNA (such as cfDNA) or RNA (such as cfRNA) in the sample can be from cancer cell(s).

[0063] As indicated above, a sample analyzed in methods of the present invention, in certain illustrative embodiments, is a blood sample, or a fraction thereof. Methods and compositions provided herein, in certain embodiments, are specially adapted for amplifying DNA fragments, especially tumor DNA fragments that are found in circulating tumor DNA (ctDNA). Such fragments are typically about 160 nucleotides in length.

[0064] It is known in the art that cell-free nucleic acid (cfNA), e.g cfDNA, can be released into the circulation via various forms of cell death such as apoptosis, necrosis, autophagy and necroptosis. The cfDNA, is fragmented and the size distribution of the fragments varies from 150-350 bp to > 10000 bp. (see Kalnina et al. World J Gastroenterol. 2015 Nov 7; 21(41): 11636-11653). For example the size distributions of plasma DNA fragments in hepatocellular carcinoma (HCC) patients spanned a range of 100-220 bp in length with a peak in count frequency at about 166bp and the highest tumor DNA concentration in fragments of 150-180 bp in length (see: Jiang et al. Proc Natl Acad Sci USA 112:E1317-E1325).

[0065] In an illustrative embodiment the circulating tumor DNA (ctDNA) is isolated from blood using EDTA-2Na tube after removal of cellular debris and platelets by centrifugation. The plasma samples can be stored at -80oC until the DNA is extracted using, for example, QIAamp DNA Mini Kit (Qiagen, Hilden, Germany), (e.g. Hamakawa et al., Br J Cancer. 2015; 112:352-356). Hamakava et al. reported median concentration of extracted cell free DNA of all samples 43.1 ng per ml plasma (range 9.5-1338 ng ml/) and a mutant fraction range of 0.001-77.8%, with a median of 0.90%.

[0066] Methods provided herein are especially effective for samples where the copy number variation is present in a small percentage of nucleic acids that are from, or are derived from the same chromosomal region exhibiting the copy number variation. That is, samples where the copy number variation (CNV) is present for less than 20, 15, or 10% of the nucleic acids in the sample that are derived from the chromosomal region with the CNV. For example, ctDNA present in less than 20%, 15%, 10% or 5%, 4%, or 3% of a cfDNA sample, are illustrative embodiments. In other embodiments, ctDNA is present in between 0.5% or 1% of a cfDNA sample on the low end of the range and 20%, 15%, 10% or 5%, 4%, or 3% of a cfDNA sample on the high end of the range. In other illustrative embodiments, the sample has an average allelic imbalance of 20% or less, 15% or less, or 10% or less, or an average allelic imbalance of 0.45%, 0.5%, 1%, 2%, 3% or 4% on the low end of the range, and 4%, 5%, 10%, 12.5%, 15%, or 20% on the high end of the range.

[0067] In certain illustrative embodiments the sample is a tumor. Methods are known in the art for isolating nucleic acid from a tumor and for creating a nucleic acid library from such a DNA sample given the teachings here. Furthermore, given the teachings herein, a skilled artisan will recognize how to create a nucleic acid library appropriate for the methods herein from other samples such as other liquid samples where the DNA is free floating in addition to ctDNA samples.

Sample Preparation

[0068] Methods of the present invention in certain embodiments, typically include a step of generating and amplifying a nucleic acid library from the sample (i.e. library preparation). The nucleic acids from the sample during the library preparation step can have ligation adapters, often referred to as library tags or ligation adaptor tags (LTs), appended, where the ligation adapters contain a universal priming sequence, followed by a universal amplification. In an embodiment, this can be done using a standard protocol designed to create sequencing libraries after fragmentation. In an embodiment, the DNA sample can be blunt ended, and then an A can be added at the 3' end. A Y-adaptor with a T-overhang can be added and ligated. In some embodiments, other sticky ends can be used other than an A or T overhang. In some embodiments, other adaptors can be added, for example looped ligation adaptors.

[0069] Primer tails can improve the detection of fragmented DNA from universally tagged libraries. If the library tag and the primer-tails contain a homologous sequence, hybridization can be improved (for example, melting temperature (Tm) is lowered) and primers can be extended if only a portion of the primer target sequence is in the sample DNA fragment. In some embodiments, 13 or more target specific base pairs can be used. In some embodiments, 10 to 12 target specific base pairs can be used. In some embodiments, 8 to 9 target specific base pairs can be used. In some embodiments, 6 to 7 target specific base pairs can be used.

[0070] In one embodiment, libraries are generated from the samples above by ligating adaptors to the ends of DNA fragments in the samples, or to the ends of DNA fragments generated from DNA isolated from the samples. The adaptors in certain embodiments, include regions that are specifically designed to bind to downstream primers used in a sequencing workflow, especially a next generation sequencing workflow, and/or include regions that can be used for universal clonal amplification. The fragments can then be amplified using PCR, using standard conditions and protocols, including, for example, the following non-limiting exemplary protocol: 95°C, 2 min; 15 x [95°C, 20 sec, 55°C, 20 sec, 68°C, 20 sec], 68°C 2 min, 4°C hold.

[0071] Many kits and methods are known in the art for generation of libraries of nucleic acids that include universal primer binding sites for subsequent amplification, for example clonal amplification, and for subsequence sequencing. To help facilitate ligation of adapters library preparation and amplification can include end repair and adenylation (i.e. A-tailing). Kits especially adapted for preparing libraries from small nucleic acid fragments, especially circulating free DNA, can be useful for practicing methods provided herein. For example, the NEXTflex Cell Free kits available from Bioo Scientific or the Natera Library Prep Kit (available from Natera, Inc. San Carlos, CA). However, such kits would typically be modified to include adaptors that are customized for the amplification and sequencing steps of the methods provided herein. Adaptor ligation can be performed using commercially available kits such as the ligation kit found in the AGILENT SURESELECT kit (Agilent, CA).

Reaction Mixtures

[0072] A number of the embodiments provided herein, including, for example, methods for determining ploidy in a ctDNA sample, include a step of receiving sequencing data for amplicons spanning each loci of a plurality or set of polymorphic loci. Such methods in illustrative embodiments, can further include an amplification step and/or a sequencing step (Sometimes referred to herein as a "ctDNA amplification/sequencing workflow) whose output is received according to a method of certain embodiments of the invention. In an illustrative example, a ctDNA amplification/sequencing workflow can include generating a set of amplicons by performing a multiplex amplification reaction on nucleic acids isolated from a sample of blood or a fraction thereof from an individual, such as an individual suspected of having a cancer, for example a lung cancer or an ovarian cancer, wherein each amplicon of the set of amplicons spans at least one polymorphic loci of a plurality or set of polymorphic loci, such as a SNP loci, known to be associated with cancer. The sequence of at least a portion of each amplicon of the plurality or set of amplicons can then be determined, wherein the portion includes a polymorphic loci.

[0073] Methods of the present invention, in certain embodiments, include forming an amplification reaction mixture. An amplification reaction mixture useful for the present invention includes some components known in the art for nucleic acid amplification, especially for PCR amplification. For example, the reaction mixture typically includes nucleotide triphosphates, a polymerase, magnesium, and primers, and optionally one or more template nucleic acids. The reaction mixture in certain embodiments, is formed by combining a polymerase, nucleotide triphosphates, nucleic acid fragments from a nucleic acid library generated from the sample, and a set of forward and/or reverse primers.

[0074] Such reaction mixtures may be used in numerous method embodiments provided herein, such as a method of determining copy number (i.e. ploidy) or detecting aneuoploidy or CNV, or methods for amplifying. Accordingly, in certain embodiments, provided herein is a reaction mixture that includes a population of circulating free nucleic acids from an individual, or nucleic acid fragments derived therefrom, and a pool of primers, at least some of which bind nucleic acids within the population of circulating free nucleic acids. The reaction mixture can include other components for an amplification reaction such as, but not limited to, a polymerase, nucleotide triphosphates, magnesium, and nucleic acid fragments from a nucleic acid library generated from the sample. The nucleic acid fragments can include adapter sequences, for example, for binding primers for sequencing reactions and/or universal amplification reactions, as discussed in more detail herein.

[0075] A composition that includes a set, plurality, library, or pool of primers or primer pairs can be part of numerous methods and other compositions provided herein. These methods include a step of amplifying nucleic acids from a sample, or for compositions, such compositions can be a reaction mixture. For any of these embodiments, the set, library, plurality or pool of primers or primer pairs can include between 25, 50, 100, 200, 250, 300, 400, 500, or 1000 primers or primer pairs on the low end of the range, and 100, 200, 250, 300, 400, 500, 1000, 1500, 2000, 2500, 5000, 10,000, or 25,000 primers or primer pairs on the high end of the range, that are each designed to amplify one or more polymorphic loci within a haploblock within a chromosomal region. For example, in one non-limiting embodiment, a set, library, plurality or pool of primers includes between 1000 and 10,000 primers each for amplifying an amplicon within a haploblock from a target chromosomal region that includes one or more polymorphic loci. Each primer of the set, plurality, or pool of primers binds an effective distance from one or more polymorphic loci, such as SNP loci, or a plurality of primer pairs in the set, library, plurality, or pool of primers, each define an amplicon that spans one or more polymorphic loci, such as a SNP loci.

[0076] The polymorphic loci, can be within genes known to be associated with cancer and are located within a haploblock. The haploblock includes 2, 3, 4, 5, 10, 20, 24, 50, 75, or 100 polymorphic loci on the low end and 4, 5, 10, 20, 24, 50, 75, 100, 150, 200 or 250 polymorphic loci on the high end of the range. One or more pools of primers can form a set of primer pools, which can include 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, or 25 pools of primers on the low end of the range, and 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 50, or 100 pools of primers on the high end of the range, that are used to form a set of reaction mixtures that can include identical amplification components except for the pool of primers. For example, a set of primers can include between 10 and 100 primers or pairs of primers per haploblock, wherein the set of primers includes 1000 to 50,000 primers.

[0077] In certain embodiments, a composition may include a set, library, plurality, or pool of primers, that includes 25, 50, 100, 200, 250, 300, 400, 500, 1,000, 2,500, 5,000, or 10,000 primers or primer pairs on the low end of the range, and 200, 250, 300, 400, 500, 1,000, 2,500, 5,000, 10,000, 20,000, 25,000, 50,000, or 100,000 primers or primer pairs on the high end of the range, that each specifically bind to a primer binding sequence located within one or more of a plurality of haploblocks found within a chromosome region known to exhibit copy number variation (CNV) associated with a disorder or disease, wherein each haploblock comprises at least 2 of the primer binding sequences and wherein at least 75, 80, 85, 90, 95, 96, 97, 98, 99%, or all of the primer binding sequences are located within haploblocks.

[0078] Methods for amplifying a set of target nucleic acids within a chromosome or chromosome region of interest (i.e. target chromosome or chromosome region of interest) of an individual may include the following:

1. a. forming a reaction mixture that includes circulating free nucleic acids derived from a sample of blood or a fragment thereof of the individual, a polymerase and a pool of primers that includes at least 500 primers or primer pairs (or any of the primer pool examples set out above) wherein at least 50, 60, 70, 75, 80, 90, 95, 96, 97, 98, 99, or 100% of the primers or primer pairs in the reaction mixture specifically bind to a primer binding sequence located within one or more of a plurality of haploblocks found within the chromosome region, wherein the chromosome region is known to exhibit copy number variation (CNV) associated with a disorder or disease; and

2. b. subjecting the reaction mixture to amplification conditions, thereby amplifying the set of target nucleic acids.

[0079] In certain examples, each haploblock includes at least 2, 3, 4, 5, 10, 15, 20, or 25 loci that have strong linkage disequilibrium with at least 1, 2, 3, 4, 5, 10, 15, 20, or 25 other loci of the plurality of loci.

[0080] In the method for amplifying, the primer or primer pairs are designed to amplify each loci of a plurality of polymorphic (e.g. SNP) loci fthat have a strong linkage disequilibrium with at least one other polymorphic loc, within one or more of a plurality of haploblocks (such as haploblocks identified based on linkage disequilibrium data from population data using publically available analysis tools (e.g. plink). In certain examples of the method for amplifying, at least 50, 60, 70, 75, 80, 90, 95, 96, 97, 98, 99, or 100% of the loci of the plurality of loci are found within the same haploblock as at least 1, 2, 3, 4, 5, 10, 15, 20, 25 other loci of the plurality of loci. In other examples, at least 50, 60, 70, 75, 80, 90, 95, 96, 97, 98, 99, or 100% of the loci of the plurality of loci have strong linkage disequilibrium with at least 1, 2, 3, 4, 5, 10, 15, 20, 25 other loci of the plurality of loci.

[0081] The size of a target chromosome region can affect the number of polymorphic loci and haploblocks selected for analysis. As illustrated in the Examples herein, for ovarian cancer using target chromosome regions greater than 50 Mb, in illustrative embodiments, haploblocks with at least 10 polymorphic loci (e.g. 10, 15, 20, or 25 polymorphic loci on the low end of the range and 15, 20, 25, 50, 100, 150, 200, 250, or 500 polymorphic loci on the high end of the range), and at least 500 or 1000 target polymorphic loci per chromosomal region and up to 1500, 2000, 2500, 5000, or 10,000 target polymorphic loci per chromosomal region, can be selected. These ranges are for finally selected polymorphic loci, which is a fraction of those available for analysis, as illustrated in the Examples herein. On the other hand, for focal chromosome regions (i.e. less than 50 Mb), minimum requirements can be relaxed. For example, for target chromosome regions that are less than 50 Mb, haploblocks with at least 2, 3, 4, or 5 polymorphic loci (e.g. 2, 3, 4, or 5 polymorphic loci on the low end of the range and 10, 15, 20, 25, 50, 100, 150, 200, or 250 on the high end of the range) and at least 100, 200, 250, 300, 400, or 500 total polymorphic loci per focused chromosome region, can be targeted. In some embodiments, depending on total number of SNPs desired for a chromosomal region, SNPs within haploblocks can be chosen starting from SNPs within the largest haploblocks. Haploblock minimum size can be determined when a minimum number of SNPs for the analysis is reached. Additional requirements or preferences for primers, loci, and amplicons can be relaxed as well, as will be apparent based on the large differences in size between large chromosome arm-level CNV in Example 1 and the focused chromosomal regions of Example 5 by comparing Example 1 and Example 5.

[0082] Exemplary primer design rules and primer selection methods are provided in Examples 1 and 5 herein. Primer designs can be generated with Primer3 (Untergrasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) "Primer3 - new capabilities and interfaces." Nucleic Acids Research 40(15):e115 and Koressaar T, Remm M (2007) "Enhancements and modifications of primer design program Primer3." Bioinformatics 23(10):1289-91) source code available at primer3.sourceforge.net). For example, primers can be designed using primer3 release 2.3.6 (Whitehead Institute for Biomedical Research, Steve Rozen (available on the Internet at primer3.sourceforge.net/releases.php) and then filtered in a reiterative fashion to check primer specificity. For each candidate SNP primer3 can be used to design left and right primers (two-sided) with an amplicon length within a range (as provided elsewhere herein, e.g. 25 to 150, 25 to 125, 25 to 100, or 50 to 75 bp) and a target melting temperature range and target temperature, for example between 50-65°C or 53-60°C. A skilled artisan will understand that target Tm ranges can be changed depending on specific amplification temperatures (e.g. annealing temperature). Primer3 can be configured to use the SantaLucia salt correction and melting temperature formulae (SantaLucia JR (1998) "A unified view of polymer, dumbbell and oligonucleotide DNA nearest-neighbor thermodynamics", Proc Natl Acad Sci 95:1460-65).

[0083] Primer specificities can be determined using the BLASTn program from the ncbi-blast-2.2.29+ package. The task option "blastn-short" can be used to map the primers against hg19 human genome. Primer designs can be determined as "specific" if the primer has less than 100 hits to the genome and the top hit is the target complementary primer binding region of the genome and is at least two scores higher than other hits (score is defined by BLASTn program). This can be done in order to have a unique hit to the genome and to not have many other hits throughout the genome.

[0084] The final selected primers can be visualized in IGV (James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24-26 (2011)) and UCSC browser (Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002 Jun;12(6):996-1006) using bed files and coverage maps for validation.

[0085] A method for selecting a plurality or set of primers for determining ploidy of a chromosomal region in a sample of an individual, or a method for selecting a primer pool for determining ploidy of a chromosomal region in a sample of an individual, or a method for selecting a plurality or set of amplicons for determining ploidy of a chromosomal region in a sample of an individual, can include the following:

1. a. identifying target chromosomal regions, wherein the target chromosomal regions are known to exhibit aneuploidy associated with a disease or disorder;
2. b. identifying target polymorphic loci within the target chromosomal regions;
3. c. identifying candidate primers for amplifying the target polymorphic loci;
4. d. filtering the candidate primers such that at least a minimum percent (e.g. 90%) of the candidate primers, and in illustrative embodiments 100% of the candidate primers bind to target loci within one of a plurality of known haploblocks; and

5. e. selecting compatible primers from the candidate primers, thereby selecting the primer pool for determining ploidy.

[0086] Such methods are exemplified in Example 1 and Example 5 herein where the target disease or disorder is cancer, and in particular ovarian cancer (Example 1) and lung cancer (Example 5). Illustrative teachings for all of these steps are found in these examples. Details provided herein for the above steps, provide embodiments that can be used in any of the methods, compositions, or kits provided herein since such methods can be part of any of the methods herein, such as part of a method of determining ploidy or detecting aneuploidy, in certain embodiments.

[0087] Details regarding identifying target chromosomal regions are provided in a separate section herein. Polymorphic loci are identified, by identifying polymorphic loci (exemplified by SNPs), that are found in, and preferably are found throughout specific genes known to exhibit CNV in a disease or disorder of interest (e.g. cancer-related genes). In preferred embodiments, even for target focused chromosomal regions, at least 1,000 SNPs are identified per target region. However, for such focused chromosomal regions involved in CNV, requirements for total number of SNPs can be relaxed, such as at least 200, 250, 300, 400, or 500 SNPs. Furthermore, polymorphic loci with a minor allele frequency of at least .1 are preferred in certain embodiments, especially for chromosome regions greater than 50 Mb. However, for focused chromosome regions, an allele frequency of .01 can be used. Filtering can be employed to eliminate certain loci, if there is not sufficient evidence that a mutation in the loci recurs.

[0088] Candidate primers for amplifying the target polymorphic loci are selected using one or more or all of a number of design rules. As disclosed herein, Primer3 can be used in the primer design process. Preferably, a SNP target loci is within the first 100, 75, and most preferably 50 nucleotides (e.g. bases) of an amplicon. Therefore, primers can be selected accordingly. Primer designs compatible with massively multiplex PCR (e.g. multiplex PCR with greater than 1000 primer pairs) in one pool with deltaG higher than -4kcal/mol are selected in illustrative embodiments. In certain embodiments, primers are selected that yield amplicons that are compatible with a downstream analysis technology, such as a high throughput sequencing technology. Preferably, primer pairs are selected such that one primer pair is selected as a left and right primer for amplifying a SNP. Primers with a Tm within a range, for example, from 50C-60C or 53C-59C can be selected, in particularly embodiments, associated with an annealing temperature that is at least 2, 3, 4, 5, 6, 7, 8, 9, or 10 degrees higher than the median Tm of the primers in a primer pool, or higher than the highest Tm of the primers in the primer pool. For example, an annealing temperature of 60-65C, such as 61-63C or 62C can be selected.

[0089] The effective distance of binding of the primers can be within 1, but in preferred embodiments, is between 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 75, 100, 125, or 150 base pairs of a polymorphic loci on the low end of the range and 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 75, 100, 125, or 150 base pairs of a polymorphic loci on the high end of the range. In some embodiments, primers bind 2-5 nucleotides from a polymorphic loci. The effective range that a pair of primers spans typically includes a polymorphic loci and is typically 160 base pairs or less, and can be 150, 140, 130, 125, 100, 75, 50 or 25 base pairs or less. In other embodiments, the effective range that a pair of primers spans is 20, 25, 30, 40, 50, 60, 70, 75, 100, 110, 120, 125, 130, 140, or 150 nucleotides within a polymorphic loci on the low end of the range, and 25, 30, 40, 50, 60, 70, 75, 100, 110, 120, 125, 130, 140, or 150, 160, 170, 175, or 200 nucleotides from a polymorphic loci on the high end of the range. Amplicons formed using primer pairs may include polymorphic loci.

[0090] An important improvement provided herein, is that by selecting primers that can be used to amplify target loci within haploblocks having a minimum number of SNP loci within a chromosome region known to exhibit aneuploidy associated with a disease or disorder, as disclosed in more detail herein, methods for determining ploidy and detecting CNV are more robust to imperfect haplotyping. Therefore, candidate primers are filtered such that at least 75, 80, 85, 90, 95, 96, 97, 98, 99 or in certain particularly illustrative embodiments 100% of the candidate primers bind to target loci within one of a plurality or set of haploblocks within the target chromosome region.

[0091] Further improvements in selecting polymorphic loci within haploblocks within target regions known to exhibit aneuploidy associated with a disease or disorder, can be input, used, and/or included in any of the disclosed methods. For example, a plurality, pool and/or set of primers includes at least 250, 300, 400, 500, or 1000 primers, and less than 100, 75, 50, 25, 10, 5, 4, 3, 2, or 1 of the primers of the plurality, pool, and/or set of primers each bind to a different target binding site that is not found in a haploblock within a target chromosome region associated with a disease or disorder, and in further exemplary embodiments, not found in a haploblock with at least 2, 3, 4, 5, or 10 polymorphic loci. Accordingly, in certain embodiments, a plurality, pool and/or set of primers includes at least 250, 300, 400, 500, or 1000 primers, and 75, 80, 85, 90, 95, 96, 97, 98, 99 or 100% of the primers of the plurality, pool, and/or set of primers each bind to a different primer binding site that is found within one of a plurality of haploblocks within a target chromosome region associated with a disease or disorder, or that binds within 2, 3, 4, 5, 6, 7, 8, 9, or 10 nucleotides from a polymorphic loci that is found within a haploblock and in further exemplary embodiments, a haploblock with at least 2, 3, 4, 5, 10 polymorphic loci. Amplicons generated and/or analyzed in methods provided herein may include amplicons that map to the human genome and amplicons that do not map to the human genome, for example because that are formed by non-specific reactions. A plurality, pool and/or set of primers includes at least 250, 300, 400, 500, or 1000 amplicons, and at least 75, 80, 85, 90, 95, 96, 97, 98, 99 or 100% of total amplicons generated or input into a method provided herein that map to a human genome, may be complementary to nucleic acid segments found within haploblocks, and in especially illustrative embodiments, haploblocks that include at least 2, 3, 4, 5, 6, 7, 8, 9, or 10 polymorphic loci.

[0092] A reaction mixture or a set of reaction mixtures or primer pools, may include a plurality, set, or library of primers or primer pairs, such as primers selected from a library of candidate primers using any of the disclosed methods. The plurality, set, or library in the reaction mixture may include primers or primer pairs that simultaneously hybridize (or are capable of simultaneously hybridizing) to or that simultaneously amplify (or are capable of simultaneously amplifying) between 100; 200; 500; 750; 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; or 25,000 different target loci on the low end of the range and 250; 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 target loci on the high end of the range, in one reaction volume. At least 50, 60, 70, 75, 80, 90, 95, 96, 97, 98, 99, or 100% of the target loci hybridized or amplified by the primers, may be within haploblocks, for example haploblocks having at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, or 15 polymorphic loci each. The pool, plurality, set, or library may include primers that simultaneously amplify (or are capable of simultaneously amplifying) between 100 to 500; 500 to 1,000; 1,000 to 2,000; 2,000 to 5,000; 5,000 to 7,500; 5,000 to 10,000; 5,000 to 20,000; 5,000 to 25,000; 5,000 to 30,000; 5,000 to 40,000; 5,000 to 50,000; 5,000 to 75,000; or 5,000 to 100,000 different target loci in one reaction volume, inclusive. The pool, plurality, set, or library in a reaction mixture, may include primers that simultaneously amplify (or are capable of simultaneously amplifying) between 1,000 to 100,000 different target loci in one reaction volume, such as between 1,000 to 50,000; 1,000 to 30,000; 1,000 to 20,000; 1,000 to 10,000; 2,000 to 30,000; 2,000 to 20,000; 2,000 to 10,000; 5,000 to 30,000; 5,000 to 20,000; or 5,000 to 10,000 different target loci, inclusive. In some embodiments, the pool, set, plurality, or library includes primers that simultaneously amplify (or are capable of simultaneously amplifying) the target loci in one reaction volume such that less than 60, 40, 30, 20, 10, 5, 4, 3, 2, 1, 0.5, 0.25, 0.1, or 0.5% of the amplified products are primer dimers. The amount of amplified products that are primer dimers is between 0.5 to 60%, such as between 0.1 to 40%, 0.1 to 20%, 0.25 to 20%, 0.25 to 10%, 0.5 to 20%, 0.5 to 10%, 1 to 20%, or 1 to 10%, inclusive. The primers simultaneously amplify (or are capable of simultaneously amplifying) the target loci in one reaction volume such that at least 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, or 99.5% of the amplified products are target amplicons. The amount of amplified products that are target amplicons is between 50 to 99.5%, such as between 60 to 99%, 70 to 98%, 80 to 98%, 90 to 99.5%, or 95 to 99.5%, inclusive. The primers simultaneously amplify (or are capable of simultaneously amplifying) the target loci in one reaction volume such that at least 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, or 99.5% of the targeted loci are amplified (e.g, amplified at least 5, 10, 20, 30, 50, or 100-fold compared to the amount prior to amplification). The amount target loci that are amplified (e.g, amplified at least 5, 10, 20, 30, 50, or 100-fold compared to the amount prior to amplification) is between 50 to 99.5%, such as between 60 to 99%, 70 to 98%, 80 to 99%, 90 to 99.5%, 95 to 99.9%, or 98 to 99.99% inclusive. 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 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. 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 individual primers that each hybridize to a different target locus, wherein the individual primers are not part of primer pairs.

[0093] The concentration of each primer may be less than 100, 75, 50, 25, 20, 10, 5, 2, or 1 nM, or less than 500, 100, 10, or 1 uM. The concentration of each primer may be between 1 uM to 100 nM, such as between 1 uM to 1 nM, 1 to 75 nM, 2 to 50 nM or 5 to 50 nM, inclusive. The GC content of the primers may be between 30 to 80%, such as between 40 to 70%, or 50 to 60%, inclusive. The range of GC content of the primers may be less than 30, 20, 10, or 5%. The range of GC content of the primers may be between 5 to 30%, such as 5 to 20% or 5 to 10%, inclusive. The melting temperature (T_m) of the test primers may be between 40 to 80 °C, such as 50 to 70 °C, 55 to 65 °C, or 57 to 60.5 °C, inclusive. The T_m may be calculated using the Primer3 program (libprimer3 release 2.2.3) using the built-in SantaLucia parameters (the world wide web at primer3.sourceforge.net) (SantaLucia JR (1998) "A unified view of polymer, dumbbell and oligonucleotide DNA nearest-neighbor thermodynamics", Proc Natl Acad Sci 95:1460-65). The range of melting temperature of the primers may be less than 15, 10, 5, 3, or 1 °C. The range of melting temperature of the primers may be between 1 to 15 °C, such as between 1 to 10 °C, 1 to 5 °C, or 1 to 3 °C, inclusive. The length of the primers may be between 15 to 100 nucleotides, such as between 15 to 75 nucleotides, 15 to 40 nucleotides, 17 to 35 nucleotides, 18 to 30 nucleotides, or 20 to 65 nucleotides, inclusive. The range of the length of the primers may be less than 50, 40, 30, 20, 10, or 5 nucleotides. The range of the length of the primers may be between 5 to 50 nucleotides, such as 5 to 40 nucleotides, 5 to 20 nucleotides, or 5 to 10 nucleotides, inclusive. The length of the target amplicons may be between 50 and 100 nucleotides, such as between 60 and 80 nucleotides, or 60 to 75 nucleotides, inclusive. The range of the length of the target amplicons may be less than 50, 25, 15, 10, or 5 nucleotides. The range of the length of the target amplicons may be between 5 to 50 nucleotides, such as 5 to 25 nucleotides, 5 to 15 nucleotides, or 5 to 10 nucleotides, inclusive. The set, plurality, pool, or library may not comprise a microarray. The set, plurality, pool, or library may comprises a microarray.

[0094] Compositions, kits, and methods provided herein for selecting a set of primers or primer pools, for determining ploidy, for detecting aneuploidy such as CNV, for detecting circulating tumor DNA, provided herein, may include the following:

at least 1000 amplicons are formed and/or input, and wthe amplicons represent at least 95% of total amplicons that map to a human genome.

at least 1000 amplicons are formed and/or input, and represent at least 99% of total amplicons that map to a human genome.

at least 1000 amplicons are formed and/or input, and represent all of the total amplicons that map to a human genome.

at least 500 primers or primer pairs are selected and/or used, wherein at least 95% of the primer or primer pairs that specifically bind to a nucleic acid in the circulating free nucleic acids and/or or that specifically bind to a genome of the individual, bind to a haploblock of a plurality or set of haploblocks, wherein the plurality of haploblocks are found within a chromosome region known to exhibit copy number variation (CNV) associated with a disorder or disease.

at least 10 polymorphic loci and at least 10 candidate primers are identified for each haploblock, and wherein at least 1000 candidate primers are identified for the primer pool and/or on a target chromosome region.

at least 10 candidate primer pairs are identified for each segment and optionally a maximum of 100 polymorphic loci and primer pairs are identified for each segment.

candidate primers are selected such that their 3' end is between 2 and 5 nucleotides away from a polymorphic loci of interest.

candidate primers are selected that form a primer pair for amplifying a segment between 50 and 75 nucleotides in length, wherein the primers are between 18 and 30 nucleotides in length and having a Tm between 50 and 60C.

candidate primers have a GC content between 30 and 70%.

polymorphic loci have a minor allele frequency of at least 10%.

[0095] The disease or disorder that the compositions and methods provided herein relate to, can include any disease or disorder correlated to allelic imbalance, copy number variation, or ploidy, especially where samples that can be used to detect, monitor, or diagnose such disease or disorder include a relatively small percentage of the total nucleic acids in a nucleic acid sample (for non-limiting example, between 1% and 25%), as set out in detail herein. For example, the disease or disorder in illustrative embodiments, is cancer, especially cancers known to involve a relative high percentage of CNVs in cancerous cells and a relatively high percentage of ctDNA.

[0096] The chromosome region may be all or a part of a chromosome known to be associated with a developmental disorder in non-invasive prenatal testing. Accordingly, the method may involve determining from a plasma sample of a mother, whether a fetus has one or more of the following conditions: cystic fibrosis, Huntington's disease, Fragile X, thallasemia, muscular dystrophy (such as Duchenne's muscular dystrophy), Alzheimer, Fanconi Anemia, Gaucher Disease, Mucolipidosis IV, Niemann-Pick Disease, Tay-Sachs disease, Sickle cell anemia, Parkinson disease, Torsion Dystonia, and cancer. In some embodiments, a target chromosome is one or more chromosomes taken from the group consisting of chromosomes 13, 18, 21, X, and Y. A fetal haplotype may be determined for all of the fetal chromosomes.

[0097] After the reaction mixture is formed it is subjected to amplification conditions to generate a set of amplicons each comprising at least one polymorphic loci of a plurality of polymorphic loci located within haploblocks, preferably known to be associated with cancer. Amplification (e.g. temperature cycling) conditions for PCR are well known in the art. The methods provided herein can include any PCR cycling conditions that result in amplification of target nucleic acids such as target nucleic acids from a library. Non-limiting exemplary cycling conditions are provided in the Examples section herein.

[0098] There are many workflows that are possible when conducting PCR; some workflows typical to the methods disclosed herein are provided herein. The steps outlined herein are not meant to exclude other possible steps nor does it imply that any of the steps described herein are required for the method to work properly. A large number of parameter variations or other modifications are known in the literature.

[0099] Following amplification (whether as part of a method of the invention or as a separate step performed outside of a method of the invention), in methods provided herein for determining ploidy that include a step of determining the sequence of an amplicon and/or haploblock, the sequence is determined for at least a portion of each amplicon of a plurality or set of amplicons, wherein the sequenced portion includes a polymorphic loci. In illustrative embodiments, the sequencing data that is generated and that is received in certain embodiments of methods provided herein, includes sequencing data that maps to the genome of the individual whose ploidy is being determined, such as the human genome, and optionally sequencing data that does not map to the genome of the individual (e.g. human genome), such as from non-specific amplicons (e.g. primer dimers). Amplicons may be within haploblocks that map to the human genome, since as discussed herein, primers may be selected to amplify polymorphic loci within haploblocks. Accordingly, over 75, 80, 90, 95, 98, 99, 99.5, 99.9, or 100% of the sequencing data generated in a method for determining ploidy herein, may map to the human genome, and over 75, 80, 90, 95, 98, 99, 99.5, 99.9, or 100% of the sequencing data that maps to the human genome is from polymorphic loci within haploblocks. The haploblocks, in certain examples, are segments that include at least 5, 10, 15, 20, 25, 50, or 100 polymorphic loci on the low end of the range, and 10, 15, 20, 25, 50, 100, 200, or 250 polymorphic loci on the high end, at least 95% of which exhibit strong linkage disequilibrim with a neighbor loci. Further disclosure regarding the size in nucleotide length and number of polymorphic loci within haplotypes are provided in other sections herein. It will be understood that the fact that at least 75% and up to 100% of sequencing data in a sequencing reaction that maps to a genome is from within haploblocks, is an important advancement over prior methods for determining ploidy using allele data from polymorphic sites, that did not utilize primer selection for targeted amplification for ploidy determination, especially from cfDNA, before sequencing, that focused on primers that amplify across polymorphic loci found within haploblocks. By selecting a primer pool that amplifies across polymorphic loci within haploblocks, methods for ploidy determination that utilize allele counts at polymorphic loci, become more robust to haplotype determination, such that the methods yield improved results when imperfect haplotype data is used.

[0100] In the method provided herein, the nucleic acid sequence of at least a portion of a nucleic acid segment that includes a polymorphic loci, and in illustrative examples the entire sequence of an amplicon, is determined. Methods for determining the sequence of an amplicon are known in the art. Any of the sequencing methods known in the art, e.g. Sanger sequencing, can be used for such sequence determination. In illustrative embodiments high throughput next-generation sequencing techniques (also referred to herein as massively parallel sequencing techniques) such as, but not limited to, those employed in MYSEQ

(Illumina), HISEQ (Illumina), ION TORRENT (Life Technologies), GENOME ANALYZER ILX (Illumina), GS FLEX+ (Roche 454), can be used for sequencing the amplicons produced by the methods provided herein. In addition, the sequence of a plurality of polymorphic loci can be determined using microarrays.

[0101] In some embodiments, the amplified products are detected using an array, such as an array especially a microarray with probes to one or more chromosomes of interest (e.g., chromosome 13, 18, 21, X, Y, or any combination thereof, or chromosome regions associated with cancer). It will be understood for example, that a commercially available SNP detection microarray could be used such as, for example, the Illumina (San Diego, CA) GoldenGate, DASL, Infinium, or CytoSNP-12 genotyping assay, or a SNP detection microarray product from Affymetrix, such as the OncoScan microarray. In some embodiments, phased genetic data for one or both biological parents of the embryo or fetus is used to increase the accuracy of analysis of array data from a single cell.

[0102] In some embodiments involving sequencing, the depth of read is the number of sequencing reads that map to a given locus. The depth of read can be normalized over the total number of reads. In some embodiments for depth of read of a sample, the depth of read is the average depth of read over the targeted loci. In some embodiments for the depth of read of a locus, the depth of read is the number of reads measured by the sequencer mapping to that locus. In general, the greater the depth of read of a locus, the closer the ratio of alleles at the locus tend to be to the ratio of alleles in the original sample of DNA. Depth of read can be expressed in variety of different ways, including but not limited to the percentage or proportion. Thus, for example in a highly parallel DNA sequencer such as an Illumina HISEQ, which, e.g., produces a sequence of 1 million clones, the sequencing of one locus 3,000 times results in a depth of read of 3,000 reads at that locus. The proportion of reads at that locus is 3,000 divided by 1 million total reads, or 0.3% of the total reads.

[0103] Allelic data is obtained, wherein the allelic data includes quantitative measurement(s) indicative of the number of copies of a specific allele of a polymorphic locus. In some embodiments, the allelic data includes quantitative measurement(s) indicative of the number of copies of each of the alleles observed at a polymorphic locus. Typically, quantitative measurements are obtained for all possible alleles of the polymorphic locus of interest. For example, any of the methods discussed in the preceding paragraphs for determining the allele for a SNP locus, such as for example, microarrays, qPCR, DNA sequencing, such as high throughput DNA sequencing, can be used to generate quantitative measurements of the number of copies of a specific allele of a polymorphic locus. This quantitative measurement is referred to herein as allelic frequency data or measured genetic allelic data. Methods using allelic data are sometimes referred to as quantitative allelic methods; this is in contrast to quantitative methods which exclusively use quantitative data from non-polymorphic loci, or from polymorphic loci but without regard to allelic identity. When the allelic data is measured using high-throughput sequencing, the allelic data typically include the number of reads of each allele mapping to the locus of interest.

[0104] In some embodiments obtaining genetic data includes (i) acquiring DNA sequence information by laboratory techniques, e.g., by the use of an automated high throughput DNA sequencer, or (ii) acquiring information that had been previously obtained by laboratory techniques, wherein the information is electronically transmitted, e.g., by a computer over the internet or by electronic transfer from the sequencing device.

[0105] High throughput genetic sequencers are amenable to the use of barcoding (i.e., sample tagging with distinctive nucleic acid sequences) so as to identify specific samples from individuals thereby permitting the simultaneous analysis of multiple samples in a single run of the DNA sequencer. The number of times a given region of the genome in a library preparation (or other nucleic preparation of interest) is sequenced (number of reads) will be proportional to the number of copies of that sequence in the genome of interest. Biases in amplification efficiency can be taken into account in such quantitative determination.

[0106] Further details regarding methods of amplification that can be used in a ctDNA amplification/sequencing workflow to determine ploidy for use in methods of the invention are provided in other sections of this specification.

Target Chromosome Regions

[0107] Target regions of a gene of interest known to exhibit aneuploidy associated with a disease or disorder are first identified in illustrative embodiments. Non-limiting exemplary methods for identifying such target regions are provided herein for identifying target chromosomal regions associated with cancer and CNV. Although the examples are set out in the context of lung cancer (Example 5) and ovarian cancer, a skilled artisan will understand that such methods can be applied to any cancer where CNV is involved. In some embodiments, the selection of the CNV regions into gain/loss enriched regions can be based on selection of CNV recurrence. In one embodiment, the selection was based on 453 ovarian patient profiles in the TCGA Ovarian Cancer Cohort. As illustrated in FIG. 1.

[0108] In some embodiments, the selection of the CNV regions into gain/loss enriched regions can be based on selection of CNV recurrence. In one embodiment, the selection was based on 453 ovarian patient profiles in the TCGA Ovarian Cancer Cohort. As illustrated in FIG. 1, three regions, Regions 1-3 were identified on chromosome 8 as regions having CNVs within 50% of peak recurrence. Regions 1 and 2 were gain regions and were split according to recurrence profile and Region 3 was a loss region. Regions 1 and 2 were split from Region 3 to maximize partitioning of amplifications versus deletions. Further, incorporated were reported amplifications and deletions identified through significance testing by TCGA, arm-level and focal events (focal events represented by vertical lines). As used herein an "arm-level" can be a CNV that spans a chromosome arm p or q. As used herein a "focal event" can be a CNV that spans a region smaller than an arm-level event. The regions were validated by interrogating COSMIC's CNV calls for the same samples (COSMIC was more conservative for calling deletions).

[0109] The identified gain/loss enriched regions were then prioritized by candidate regions based on the number of ovarian patients that have a CNV in the region. FIG. 2 illustrates the identification of 14 regions, nine gain regions and five loss region. There is a correlation in that CNV events co-occur within and between patients (FIGS. 3A-3B). The table represents the pairwise Pearson correlation between the 14 regions based on the presence or absence of events across those patients captured by each of the 14 regions. Reported values are Pearson R-squared values. Bold entries indicate positive correlation, boxed entries indicate correlation < 0.1. The genomic coordinates of each of the locations is reported in GRCh37 coordinates.

[0110] If the deletion regions are excluded, patient coverage is reduced by 7-8% while if the regions are ranked using COSMIC call, nine of the top ranking regions are amplifications. When removing deletions at most two copies can be lost as deletions have a limited signal whereas amplifications can have more than two copies gained. Thus, amplifications can provide a better signal for CNV calling. Chromosome number and locations are listed on the x-axis and cumulative patient coverage is listed on the y-axis.

[0111] In some embodiments, the selection of the CNV chromosome target regions can be based on CNV recurrence analysis in a population of cancer patients. In one embodiment, the selection was based on 453 ovarian patient profiles in the TCGA Ovarian Cancer Cohort. As illustrated in FIG. 1, three regions, Regions 1-3 were identified on chromosome 8 as regions having CNVs within 50% peak recurrence. Regions 1 and 2 were gain regions and were split according to recurrence profile and Region 3 was a loss region. Regions 1 and 2 were split from Region 3 to maximize portioning of amplifications versus deletions. Further, incorporated were reported amplifications and deletions identified through significance testing by TCGA, arm-level and focal events (vertical lines). The regions were validated by interrogating COSMIC's CNV calls for the same samples (COSMIC was more conservative for calling deletions).

[0112] In some embodiments, a pooling algorithm is created for analyzing the haploblock data. The chromosomal segments/regions used to form a haploblock can have candidate SNPs selected from the 1000GP database with MAF > 10%. These blocks were identified using the 1000GP reference panel. PCR assays for the selected SNPs are designed in a reiterative process to allow for massive multiplexing PCR. Assays within small haploblocks, i.e., haploblocks having <10 CNVs, are filterd. The resulting optimized set of non-interactive assays are selected and can be further optimized by evaluating: The total number of patients with CNVs covering at least 50% of the region; the recurrence profile of each patient; the size of the haploblock; the MAF, population diversity and heterozygosity rate for each SNP; the type of mutation, transversion or transition; and the length of the amplicon, Tm and GC-content.

[0113] In some embodiments, an in silico simulation of the use of designed assays can be run to refine use of haploblocks for detection. To illustrate, an in silico experiment simulates use of HCC1954 and HCC2218 in a titration experiment using the blocks from the described design criteria above. It is assumed that there is perfect information within the blocks and no information between the blocks. Blocks of a minimum size of CNVs are tested with sizes of 1, 10, 15 and 20 CNVs. It was found that performance stabilizes around a minimum block size of 10-15 as too many false positives resulted from not having a minimum block requirement. Using a minimum block size of 10 it was found that performance was similar to perfect haplotypes in regions with >1000 SNPs (down to 0.5% allelic imbalance detection with some false positives). A poor region (having approximately 300 SNPs in blocks) had detection around 2% allelic imbalance. Tables 1A-1B illustrate in silico results in single pools for the designed regions.

Table 1A: *covering 436 patients out of 453

Chrom	Start	End	Patients	Number of Assays
8*	115,298,000	145,233,000	173	1451
3*	166,356,000	180,256,000	108	1364
8*	100,758,000	115,298,000	101	1490
8*	617,000	37,343,000	99	1452
19*	28,240,000	33,433,000	82	1376
20*	29,369,569	63,025,520	82	1483
20*	1	26,369,569	67	1568
12*	18,959,000	29,050,000	65	1186
19*	34,341,000	40,857,000	55	1225
19	12,042,000	17,796,000	54	903
16*	60,437,000	89,380,000	50	1480
17	25,800,001	31,800,000	30	841
22*	42,378,000	49,332,000	21	1574
17	10,700,001	16,000,000	16	535

Table 1B: *covering 436 patients out of 453

Chrom	Expected Number of hets	SNPs in blocks > 10	Yield	Longest block	Haplotyping error rate
8*	568	1381	95%	66	3.00%
3*	496	1202	88%	49	3.60%
8*	554	1464	98%	68	2.40%
8*	538	1406	97%	65	3.20%
19*	520	1237	90%	64	2.10%
20*	553	1420	96%	65	3.10%
20*	593	1537	98%	93	3.00%
12*	414	1035	87%	62	3.20%
19*	455	1105	90%	57	2.70%
19	330	731	81%	37	3.80%
16*	534	1398	94%	53	3.20%

Chrom	Expected Number of hets	SNPs in blocks > 10	Yield	Longest block	Haplotyping error rate
17	321	749	89%	35	2.80%
22*	612	1168	74%	66	3.40%
17	195	429	80%	21	5.30%

[0114] In some embodiments, the regions can be enriched for gain (amplification) or loss (deletion). FIGS. 4A-4H illustrate gain/loss enriched regions for selected chromosomes. The graphs illustrate the gain/loss enriched regions as a lined box: above the x-axis is a gain, below the x-axis is a loss, and the solid dashed line below the x-axis indicates the centromere position between the chromosome arms. Specifically known cancer genes are also identified. There were 15 candidate gain/loss enriched regions identified, nine regions were enriched for amplifications (gains) and six regions were enriched for deletions (loss). A deletion was included because the region spans cancer census genes and/or was reported to distinguish between ovarian subtypes. Chromosome number and locations are listed on the x-axis and cumulative patient coverage is listed on the y-axis. FIG. 4A, chromosome 3, one gain enriched region is illustrated, the region spans PIK3CA gene. FIG. 4B, chromosome 8, two gain and one loss enriched regions are illustrated, the region spans MYC gene. FIG. 4C, chromosome 12, one gain enriched region is illustrated, the region spans KRAS gene. FIG. 4D, chromosome 13, one loss enriched region is illustrated, the region spans RB1 gene, whose CNV status in patients has been reported to stratify clear cell and serous and ovarian cancer subtypes and GISTIC focal event inference. FIG. 4E, chromosome 16, one loss enriched region is illustrated, the region spans CDH1 gene. FIG. 4F, chromosome 17, two loss enriched regions are illustrated, the region spans MAP2K4 AND NF1 genes. Chromosome 17 was included based on GISTIC arm-level inference. FIG. 4G, chromosome 19, three gain enriched regions are illustrated, the region spans CCNE1, which is diagnostic of poor patient survival, and AKT2 genes. FIG. 4H, chromosome 20, two gain enriched regions are illustrated. Inclusion of chromosome 20 was based on GISTIC arm-level inference. GISTIC refers to an algorithm that infers the statistical significance of either gain or loss recurrence within a patient cohort. GISTIC was applied to the TCGA data and published in the TCGA Ovarian Cancer publication ("Integrated genomic analysis of ovarian carcinoma" Nature 474:609-616. 2011).

[0115] Chromosome regions exhibiting CNV can be either arm-level CNVs or focal (<50Mb) events, and methods provided herein can analyze either type of CNV. Example 1 provides an example of arm-length CNV detection. Example 5 provides an example of focal CNV detection. Accordingly, in certain embodiments, the target chromosome region is greater than 50 Mb and in other embodiments, the target chromosome region 50 Mb or less or is less than 50 Mb, or for example 10 Mb, 15 Mb, 20 Mb, 25 Mb, 30 Mb, 40 Mb, on the low end of the range and 15 Mb, 20 Mb, 25 Mb, 30 Mb, 40 Mb, 45 Mb, or 50 Mb on the high end of the range.

[0116] George et. Al. 2015 provides an algorithm for copy number analyses called CGRAS, which uses Rank sums and smoothing procedures. Statistics of smoothed rank sum profiles are computed to determine significant copy-number alterations. Additional processes can then be applied, such as those shown in Example 5, to assist in a final determination of target chromosome region(s). Chromosome regions that show CNV in at least 50, 60, 70, 80, or 90% of samples from individuals with a target disease or disorder are selected, in illustrative embodiments. In embodiments, chromosome regions that include driver genes are selected.

[0117] Target regions of the nucleic acid library generated from DNA isolated from the sample, especially a circulating free DNA sample for the methods of the present invention, are then amplified. For this amplification, a series of primers or primer pairs, which can include between 5, 10, 15, 20, 25, 50, 100, 125, 150, 250, 500, 1000, 2500, 5000, 10,000, 20,000, 25,000, or 50,000 on the low end of the range and 15, 20, 25, 50, 100, 125, 150, 250, 500, 1000, 2500, 5000, 10,000, 20,000, 25,000, 50,000, 60,000, 75,000, or 100,000 primers on the upper end of the range, that each bind to one of a series of primer binding sites.

[0118] A plurality of chromosome regions have been identified, as illustrated in the Examples section herein, that are particularly effective when detecting, diagnosing, and/or determining an effective treatment plan or identifying an effective therapeutic for ovarian cancer (Examples 1-4; See Example 1 for target chromosome regions) and a plurality of chromosome regions have been identified that are particularly effective when detecting, diagnosing, and/or determining a effective treatment plan or identifying an effective therapeutic, for lung cancer (Example 5; lung cancer therapeutic target chromosome regions provided in Example 5). The exemplary target chromosome regions for ovarian cancer include chromosome 8 nucleotides 115,298,000 - 145,233,000, chromosome 8 nucleotides 100758000-115298000, chromosome 8 nucleotides 617000-37343000, chromosome 3 nucleotides 166356000-180256000, chromosome 22 nucleotides 42378000-49332000, chromosome 19 nucleotides 34341000-40857000, chromosome 19 nucleotides 28240000-33433000, chromosome 19 nucleotides 12042000-17796000, chromosome 16 nucleotides 60437000-89380000, chromosome 12 nucleotides 18959000-29050000, chromosome 20 nucleotides 1-26369569, chromosome 20 nucleotides 29369569-63025520, chromosome 17 nucleotides 25800001-31800000, and chromosome 17 nucleotides 10700001-16000000. Methods of the present invention, include determining or estimating a phase of a plurality of polymorphic loci within a set of chromosomes that includes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, or all 14 of the above target chromosome regions.

[0119] The exemplary target chromosome regions for lung cancer that have been identified, as illustrated in the Examples herein, are regions that are particularly well-suited for targeted therapy and include chromosome 7 nucleotides 140433813-140624564 (BRAF), chromosome 7 nucleotides 55086725-55275031 (EGFR), chromosome 17 nucleotides 37856231-37884915 (ERBB2), chromosome 8 nucleotides 38268656-38325363 (FGFR1), chromosome 12 nucleotides 25358180-25403854 (KRAS), chromosome 7 nucleotides 116312459-116438440 (MET), chromosome 8 nucleotides 128748315-128753680 (MYC), and chromosome 3 nucleotides 178866311-178952497 (PIK3CA). Methods of the present invention, include determining or estimating a phase of a plurality of polymorphic loci within a set of chromosome regions that includes 1, 2, 3, 4, 5, 6, 7, or all 8 of the above target chromosome regions.

[0120] It will be understood that the above chromosome regions identified for ovarian and lung CNV provide guideposts and that regions that include at least 50, 60, 70, 75, 80, 90, 95, 98, 99, or 100% of the contiguous nucleic acids of the above regions could be useful in the methods of the invention, or regions that include 50, 60, 70, 75, 80, 90, 95, 98, 99, or 100% of the polymorphic loci within the target chromosome regions. Accordingly, in some embodiments, methods of the invention include analyzing between: 50% - 100% of the contiguous nucleic acids of the exemplary target chromosome regions, 60% - 99% of the contiguous nucleic acids of the target chromosome regions, 65% - 95% of the contiguous nucleic acids of the target chromosome regions, 70% - 90% of the contiguous nucleic acids of the target chromosome regions, and 75% - 85% of the contiguous nucleic acids of the target chromosome regions. In some embodiments, at least 75. 80, 85, 90, 95, 98, or 99%, or all of the contiguous nucleic acids of each chromosome region of the set of chromosome regions are analyzed. In some embodiments, the target chromosome region includes 5, 10, 15, 20, 25, 50, 75, or 100% more of a chromosomal region than includes the exemplary target chromosome regions. In some embodiments, the analysis is nucleic acid sequencing of the entire region. However, in illustrative embodiments, the analyzing is determining the nucleic acid sequence of polymorphic loci within haploblocks within the chromosome regions using targeted amplification and sequencing.

Exemplary methods for determining whether ctDNA is present

[0121] Chromosomal regions may be employed in a method for determining whether circulating tumor nucleic acids from a cancer, such as an Ovarian cancer or lung cancer, are present in a liquid sample from an individual, comprising: analyzing the sample to determine a ploidy at a plurality of chromosome regions in the individual, wherein the analyzing comprises separately analyzing SNP allelic data for between 10 and 100 SNP loci within a set of chromosome segments from each of the plurality of chromosome regions, and then combining the separate SNP allelic data to determine a segment allele for each of the set of chromosome segments, and then combining segment allelic data for segments on the same chromosome region to determine ploidy of each of the chromosome regions; and determining the level of allelic imbalance present for each chromosome region of the plurality of chromosome regions based on the ploidy determination, whereby an allelic imbalance above a cutoff value is indicative of the presence of circulating tumor nucleic acids. As illustrated in Tables 1A-1B, the number of SNPs in a chromosomal region and the number of SNP and haplotype blocks in a given chromosome region can provide information for detecting chromosomal aneuploidy.

[0122] The method may further include detecting a single nucleotide variant at a single nucleotide variance site in a set of single nucleotide variance locations, wherein detecting either an allelic imbalance equal to or greater than 0.45% or detecting the single nucleotide variant, or both, is indicative of the presence of circulating tumor nucleic acids in the sample. Accordingly, such methods have the advantage of analyzing for either or both SNVs and CNVs, to increase the performance of the test method.

[0123] The method for determining whether circulating tumor nucleic acids from an Ovarian cancer are present in the liquid sample may comprise analyzing a plurality of chromosome regions comprise at least two segments selected from the group of chromosome regions consisting of at least 70%, at least 80%, at least 85%, at least 90%, at least 95% and at least 99% of the contiguous nucleotides of the following plurality of chromosome regions: chromosome 8 nucleotides 115,298,000 - 145,233,000, chromosome 8 nucleotides 100758000-115298000, chromosome 8 nucleotides 617000-37343000, chromosome 3 nucleotides 166356000-180256000, chromosome 22 nucleotides 42378000-49332000, chromosome 19 nucleotides 34341000-40857000, chromosome 19 nucleotides 28240000-33433000, chromosome 19 nucleotides 12042000-17796000, chromosome 16 nucleotides 60437000-89380000, chromosome 12 nucleotides 18959000-29050000, chromosome 20 nucleotides 1-26369569, chromosome 20 nucleotides 29369569-63025520, chromosome 17 nucleotides 25800001-31800000, chromosome 17 nucleotides 10700001-16000000. The group of chromosome regions may consist of at least 70%, at least 80%, at least 85%, at least 90%, at least 95% and at least 99% of the contiguous nucleotides of the plurality of chromosome regions. In one embodiment, each chromosome region in the plurality of chromosome regions comprises a plurality of segments of between: 20 and 600 segments, 30 and 550 segments, 75 and 500 segments, and 100 and 350 segments.

[0124] Each chromosome region in the plurality of chromosome regions may comprise at least two chromosome regions from at least two chromosomes selected from the group consisting of chromosome 3, chromosome 8, chromosome 12, chromosome 13, chromosome 16, chromosome 19, chromosome 20, and chromosome 22. The plurality of chromosome regions may comprise at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, or 13 or all 14 segments selected from the group of chromosome regions consisting of at least the following plurality of chromosome regions: chromosome 8 nucleotides 115,298,000 - 145,233,000, chromosome 8 nucleotides 100758000-115298000, chromosome 8 nucleotides 617000-37343000, chromosome 3 nucleotides 166356000-180256000, chromosome 22 nucleotides 42378000-49332000, chromosome 19 nucleotides 34341000-40857000, chromosome 19 nucleotides 28240000-33433000, chromosome 19 nucleotides 12042000-17796000, chromosome 16 nucleotides 60437000-89380000, chromosome 12 nucleotides 18959000-29050000, chromosome 20 nucleotides 1-26369569, chromosome 20 nucleotides 29369569-63025520, chromosome 17 nucleotides 25800001-31800000, chromosome 17 nucleotides 10700001-16000000. The set of chromosome segments from each of the plurality of chromosome regions a plurality of segments may comprise between: 50% - 100% of the chromosome segments, 60% - 99% of the chromosome segments, 65% - 95% of the chromosome segments, 70% - 90% of the chromosome segments, and 75% - 85% of the chromosome segments. The analyzing may be performed using high throughput nucleic acid sequencing by determining the nucleic acid sequence of less than 10% of the nucleotides within each segment of the plurality of chromosome regions.

[0125] A method for determining whether circulating tumor nucleic acids from an Ovarian cancer are present in a liquid sample from an individual, may comprise analyzing the sample to determine a ploidy at a plurality of chromosome regions in the individual, wherein the chromosome regions comprise at least two segments that exhibit copy number variation in at least 50% of Ovarian cancer patients; and determining the level of allelic imbalance present for each chromosome region of the set of chromosome regions based on the ploidy determination, wherein an allelic imbalance equal to or greater than 0.45% for any of the chromosome regions is indicative of the presence of circulating tumor nucleic acids in the sample. I some embodiments, the analyzing comprises separately analyzing SNP allelic data for between 10 and 100 SNP loci with strong linkage disequilibrium within each segment of a set of chromosome segments from each of the plurality of chromosome regions, and then combining the separate SNP allelic data to determine a segment allele for each of the set of chromosome segments, and then combining segment allelic data for segments on the same chromosome region to determine ploidy of each of the chromosome regions. The analyzing may comprise analyzing at least two chromosome segments selected from the group of chromosome regions consisting of the following plurality of chromosome regions: chromosome 8 nucleotides 115,298,000 - 145,233,000, chromosome 8 nucleotides 100758000-115298000, chromosome 8 nucleotides 617000-37343000, chromosome 3 nucleotides 166356000-180256000, chromosome 22 nucleotides 42378000-49332000, chromosome 19 nucleotides 34341000-40857000, chromosome 19 nucleotides 28240000-33433000, chromosome 19 nucleotides 12042000-17796000, chromosome 16 nucleotides 60437000-89380000, chromosome 12 nucleotides 18959000-29050000, chromosome 20 nucleotides 1-26369569, chromosome 20 nucleotides 29369569-63025520, chromosome 17 nucleotides 25800001-31800000, and chromosome 17 nucleotides 10700001-16000000 for an average allelic imbalance indicative of a deletion of the segment.

[0126] The method for determining whether circulating tumor nucleic acids from an Ovarian cancer are present in a liquid sample from an individual, comprises detecting a single nucleotide variant at a single nucleotide variance site in a set of single nucleotide variance locations, wherein detecting either an allelic imbalance equal to or greater than 0.45% or detecting the single nucleotide variant, or both, is indicative of the presence of circulating tumor nucleic acids in the sample. The method may comprise performing the method on an Ovarian cancer control nucleic acid sample with a known average allelic imbalance ratio and the control can be a chromosomal region sample from the tumor of the individual. The analyzing of the sample may comprise performing a multiplex PCR to amplify amplicons across 1000 to 50,000 polymeric loci on the set of chromosome regions.

Target Genes

[0127] Target genes of the present invention in exemplary embodiments, are cancer-related genes. A cancer-related gene (for example, an ovarian cancer-related gene, a lung cancer-related gene or a lung SCC-related gene or a lung ADC-related gene) refers to a gene associated with an altered risk for a cancer (e.g. ovarian cancer, lung cancer or lung SCC or lung ADC, respectively) or an altered prognosis for a cancer. or a target for a cancer therapy. 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.

[0128] Exemplary polymorphisms or mutations (such as deletions or duplications) detected by methods provided herein are in one or more of the following genes: TP53, PTEN, PIK3CA, APC, EGFR, NRAS, NF2, FBXW7, ERBBs, ATAD5, KRAS, BRAF, VEGF, EGFR, HER2, ALK, p53, BRCA, BRCA1, BRCA2, SETD2, LRP1B, PBRM, SPTA1, DNMT3A, ARID1A, GRIN2A, TRRAP, STAG2, EPHA3/5/7, POLE, SYNE1, C20orf80, CSMD1, CTNNB1, ERBB2. FBXW7, KIT, MUC4, ATM, CDH1, DDX11, DDX12, DSPP, EPPK1, FAM186A, GNAS, HRNR, KRTAP4-11, MAP2K4, MLL3, NRAS, RB1, SMAD4, TTN, ABCC9, ACVR1B, ADAM29, ADAMTS19, AGAP10, AKT1, AMBN, AMPD2, ANKRD30A, ANKRD40, APOBR, AR, BIRC6, BMP2, BRAT1, BTNL8, C12orf4, C1QTNF7, C20orf186, CAPRIN2, CBWD1, CCDC30, CCDC93, CD5L, CDC27, CDC42BPA, CDH9, CDKN2A, CHD8, CHEK2, CHRNA9, CIZ1, CLSPN, CNTN6, COL14A1, CREBBP, CROCC, CTSF, CYP1A2, DCLK1, DHDDS, DHX32, DKK2, DLEC1, DNAH14, DNAH5, DNAH9, DNASE1L3, DUSP16, DYNC2H1, ECT2, EFHB, RRN3P2, TRIM49B, TUBB8P5, EPHA7, ERBB3, ERCC6, FAM21A, FAM21C, FCGBP, FGFR2, FLG2, FLT1, FOLR2, FRYL, FSCB, GAB1, GABRA4, GABRP, GH2, GOLGA6L1, GPHB5, GPR32, GPX5, GTF3C3, HECW1, HIST1H3B, HLA-A, HRAS, HS3ST1, HS6ST1, HSPD1, IDH1, JAK2, KDM5B, KIAA0528, KRT15, KRT38, KRTAP21-1, KRTAP4-5, KRTAP4-7, KRTAP5-4, KRTAP5-5, LAMA4, LATS1, LMF1, LPAR4, LPPR4, LRRFIP1, LUM, LYST, MAP2K1, MARCH1, MARCO, MB21D2, MEGF10, MMP16, MORC1, MRE11A, MTMR3, MUC12, MUC17, MUC2, MUC20, NBPF10, NBPF20, NEK1, NFE2L2, NLRP4, NOTCH2, NRK, NUP93, OBSCN, OR11H1, OR2B11, OR2M4, OR4Q3, OR5D13, OR8I2, OXSM, PIK3R1, PPP2R5C, PRAME, PRF1, PRG4, PRPF19, PTH2, PTPRC, PTPRJ, RAC1, RAD50, RBM12, RGPD3, RGS22, ROR1, RP11-671M22.1, RP13-996F3.4, RP1L1, RSBN1L, RYR3, SAMD3, SCN3A, SEC31A, SF1, SF3B1, SLC25A2, SLC44A1, SLC4A11, SMAD2, SPTA1, ST6GAL2, STK11, SZT2, TAF1L, TAX1BP1, TBP, TGFBI, TIF1, TMEM14B, TMEM74, TPTE, TRAPPC8, TRPS1, TXNDC6, USP32, UTP20, VASN, VPS72, WASH3P, WWTR1, XPO1, ZFHX4, ZMIZ1, ZNF167, ZNF436, ZNF492, ZNF598, ZRSR2, ABL1, AKT2, AKT3, ARAF, ARFRP1, ARID2, ASXL1, ATR, ATRX, AURKA, AURKB, AXL, BAP1, BARD1, BCL2, BCL2L2, BCL6, BCOR, BCORL1, BLM, BRIP1, BTK, CARD11, CBFB, CBL, CCND1, CCND2, CCND3, CCNE1, CD79A, CD79B, CDC73, CDK12, CDK4, CDK6, CDK8, CDKN1B, CDKN2B, CDKN2C, CEBPA, CHEK1, CIC, CRKL, CRLF2, CSF1R, CTCF, CTNNA1, DAXX, DDR2, DOT1L, EMSY (C11orf30), EP300, EPHA3, EPHA5, EPHB1, ERBB4, ERG, ESR1, EZH2, FAM123B (WTX), FAM46C, FANCA, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCL, FGF10, FGF14, FGF19, FGF23, FGF3, FGF4, FGF6, FGFR1, FGFR2, FGFR3, FGFR4, FLT3, FLT4, FOXL2, GATA1, GATA2, GATA3, GID4 (C17orf39), GNA11, GNA13, GNAQ, GNAS, GPR124, GSK3B, HGF, IDH1, IDH2, IGF1R, IKBKE, IKZF1, IL7R, INHBA, IRF4, IRS2, JAK1, JAK3, JUN, KAT6A (MYST3), KDM5A, KDM5C, KDM6A, KDR, KEAP1, KLHL6, MAP2K2, MAP2K4, MAP3K1, MCL1, MDM2, MDM4, MED12, MEF2B, MEN1, MET, MITF, MLH1, MLL, MLL2, MPL, MSH2, MSH6, MTOR, MUTYH, MYC, MYCL1, MYCN, MYD88, NF1, NFKBIA, NKX2-1, NOTCH1, NPM1, NRAS, NTRK1, NTRK2, NTRK3, PAK3, PALB2, PAX5, PBRM1, PDGFRA, PDGFRB, PDK1, PIK3CG, PIK3R2, PPP2R1A, PRDM1, PRKAR1A, PRKDC, PTCH1, PTPN11, RAD51, RAF1, RARA, RET, RICTOR, RNF43, RPTOR, RUNX1, SMARCA4, SMARCB1, SMO, SOCS1, SOX10, SOX2, SPEN, SPOP, SRC, STAT4, SUFU, TET2, TGFBR2, TNFAIP3, TNFRSF14, TOP1, TP53, TSC1, TSC2, TSHR, VHL, WISP3, WT1, ZNF217, ZNF703, and combinations thereof (Su et al., J Mol Diagn 2011, 13:74-84; DOI:10.1016/j.jmoldx.2010.11.010; and Abaan et al., "The Exomes of the NCI-60 Panel: A Genomic Resource for Cancer Biology and Systems Pharmacology", Cancer Research, July 15, 2013). In some embodiments, the duplication is a chromosome 1p ("Chr1p") duplication associated with breast cancer. In some embodiments, one or more polymorphisms or mutations are in BRAF, such as the V600E mutation. In some embodiments, one or more polymorphisms or mutations are in K-ras. In some embodiments, there is a combination of one or more polymorphisms or mutations in K-ras and APC. In some embodiments, there is a combination of one or more polymorphisms or mutations in K-ras and p53. In some embodiments, there is a combination of one or more polymorphisms or mutations in APC and p53. In some embodiments, there is a combination of one or more polymorphisms or mutations in K-ras, APC, and p53. In some embodiments, there is a combination of one or more polymorphisms or mutations in K-ras and EGFR. Exemplary polymorphisms or mutations are in one or more of the following microRNAs: miR-15a, miR-16-1, miR-23a, miR-23b, miR-24-1, miR-24-2, miR-27a, miR-27b, miR-29b-2, miR-29c, miR-146, miR-155, miR-221, miR-222, and miR-223 (Calin et al. "A microRNA signature associated with prognosis and progression in chronic lymphocytic leukemia." N Engl J Med 353:1793- 801, 2005,

[0129] 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.

[0130] 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.

[0131] In some embodiments, the tandem repeat is a repeat of between 2 and 60 nucleotides, such as 2 to 6, 7 to 10, 10 to 20, 20 to 30, 30 to 40, 40 to 50, or 50 to 60 nucleotides, inclusive. In some embodiments, the tandem repeat is a repeat of 2 nucleotides (dinucleotide repeat). In some embodiments, the tandem repeat is a repeat of 3 nucleotides (trinucleotide repeat).

[0132] In some embodiments, the polymorphism or mutation is prognostic. Exemplary prognostic mutations include K-ras mutations, such as K-ras mutations that are indicators of postoperative disease recurrence in colorectal cancer (Ryan et al. " A prospective study of circulating mutant KRAS2 in the serum of patients with colorectal neoplasia: strong prognostic indicator in postoperative follow up," Gut 52:101-108, 2003; and Lecomte T et al. Detection of free-circulating tumor-associated DNA in plasma of colorectal cancer patients and its association with prognosis," Int J Cancer 100:542-548, 2002.

[0133] Methods provided herein can be used to detect CNVs known to be associated with lung cancer. Exemplary lung cancer CNVs can be in one or more of the following genes: EGFR, FGFR1, FGFR2, ALK, MET, ROS1, NTRK1, RET, HER2, DDR2, PDGFRA, KRAS, NF1, BRAF, PIK3CA, MEK1, NOTCH1, MLL2, EZH2, TET2, DNMT3A, SOX2, MYC, KEAP1, CDKN2A, NRG1, TP53, LKB1, and PTEN, which have been identified in various lung cancer samples as being mutated, having increased copy numbers, or being fused to other genes and combinations thereof (Non-small-cell lung cancers: a heterogeneous set of diseases. Chen et al. Nat. Rev. Cancer. 2014 Aug 14(8):535-551). In illustrative embodiments, a method of the invention is directed to determining ploidy in an individual that is screened for, or suspected of having Ovarian cancer, and the target chromosome regions are found in the MYC, PIK3CA, CCNE1, KRAS, AKT2, CDH1, NF1, RB1, and/or MAP2K4 genes, as illustrated in Example 1. In other illustrative embodiments, a method of the invention is directed to determining ploidy in an individual that is screened for, or suspected of having lung cancer, and the target chromosome regions are found in the BRAF, EGFR, ERBB2, FGFR1, KRAS, MET, MYC and/or PIK3CA genes. Such methods can further include recommending administration or, or administering a targeted therapeutic agent, such as those identified in Example 5 herein.

Exemplary Cancers

[0134] Exemplary diseases or disorders that can be diagnosed, prognosed, stabilized, treated, or prevented using any of the methods disclosed, may include solid tumors, carcinomas, sarcomas, lymphomas, leukemias, germ cell tumors, or blastomas. The cancer may be an acute lymphoblastic leukemia, acute myeloid leukemia, adrenocortical carcinoma, AIDS-related cancer, AIDS-related lymphoma, anal cancer, appendix cancer, astrocytoma (such as childhood cerebellar or cerebral astrocytoma), basal-cell carcinoma, bile duct cancer (such as extrahepatic bile duct cancer) bladder cancer, bone tumor (such as osteosarcoma or malignant fibrous histiocytoma), brainstem glioma, brain cancer (such as cerebellar astrocytoma, cerebral astrocytoma/malignant glioma, ependymo, medulloblastoma, supratentorial primitive neuroectodermal tumors, or visual pathway and hypothalamic glioma), glioblastoma, breast cancer, bronchial adenoma or carcinoid, burkitt's lymphoma, carcinoid tumor (such as a childhood or gastrointestinal carcinoid tumor), carcinoma central nervous system lymphoma, cerebellar astrocytoma or malignant glioma (such as childhood cerebellar astrocytoma or malignant glioma), cervical cancer, childhood cancer, chronic lymphocytic leukemia, chronic myelogenous leukemia, chronic myeloproliferative disorders, colon cancer, cutaneous t-cell lymphoma, desmoplastic small round cell tumor, endometrial cancer, ependymoma, esophageal cancer, ewing's sarcoma, tumor in the ewing family of tumors, extracranial germ cell tumor (such as a childhood extracranial germ cell tumor), extragonadal germ cell tumor, eye cancer (such as intraocular melanoma or retinoblastoma eye cancer), gallbladder cancer, gastric cancer, gastrointestinal carcinoid tumor, gastrointestinal stromal tumor, germ cell tumor (such as extracranial, extragonadal, or ovarian germ cell tumor), gestational trophoblastic tumor, glioma (such as brain stem, childhood cerebral astrocytoma, or childhood visual pathway and hypothalamic glioma), gastric carcinoid, hairy cell leukemia, head and neck cancer, heart cancer, hepatocellular (liver) cancer, hodgkin lymphoma, hypopharyngeal cancer, hypothalamic and visual pathway glioma (such as childhood visual pathway glioma), islet cell carcinoma (such as endocrine or pancreas islet cell carcinoma), kaposi sarcoma, kidney cancer, laryngeal cancer, leukemia (such as acute lymphoblastic, acute myeloid, chronic lymphocytic, chronic myelogenous, or hairy cell leukemia), lip or oral cavity cancer, liposarcoma, liver cancer (such as non-small cell or small cell cancer), lung cancer, lymphoma (such as AIDS-related, burkitt, cutaneous T cell, Hodgkin, non-hodgkin, or central nervous system lymphoma), macroglobulinemia (such as waldenström macroglobulinemia, malignant fibrous histiocytoma of bone or osteosarcoma, medulloblastoma (such as childhood medulloblastoma), melanoma, merkel cell carcinoma, mesothelioma (such as adult or childhood mesothelioma), metastatic squamous neck cancer with occult, mouth cancer, multiple endocrine neoplasia syndrome (such as childhood multiple endocrine neoplasia syndrome), multiple myeloma or plasma cell neoplasm. mycosis fungoides, myelodysplastic syndrome, myelodysplastic or myeloproliferative disease, myelogenous leukemia (such as chronic myelogenous leukemia), myeloid leukemia (such as adult acute or childhood acute myeloid leukemia), myeloproliferative disorder (such as chronic myeloproliferative disorder), nasal cavity or paranasal sinus cancer, nasopharyngeal carcinoma, neuroblastoma, oral cancer, oropharyngeal cancer, osteosarcoma or malignant fibrous histiocytoma of bone, ovarian cancer, ovarian epithelial cancer, ovarian germ cell tumor, ovarian low malignant potential tumor, pancreatic cancer (such as islet cell pancreatic cancer), paranasal sinus or nasal cavity cancer, parathyroid cancer, penile cancer, pharyngeal cancer, pheochromocytoma, pineal astrocytoma, pineal germinoma. pineoblastoma or supratentorial primitive neuroectodermal tumor (such as childhood pineoblastoma or supratentorial primitive neuroectodermal tumor), pituitary adenoma, plasma cell neoplasia, pleuropulmonary blastoma, primary central nervous system lymphoma, cancer, rectal cancer, renal cell carcinoma, renal pelvis or ureter cancer (such as renal pelvis or ureter transitional cell cancer, retinoblastoma, rhabdomyosarcoma (such as childhood rhabdomyosarcoma), salivary gland cancer, sarcoma (such as sarcoma in the ewing family of tumors, Kaposi, soft tissue, or uterine sarcoma), sézary syndrome, skin cancer (such as nonmelanoma, melanoma, or merkel cell skin cancer), small intestine cancer, squamous cell carcinoma, supratentorial primitive neuroectodermal tumor (such as childhood supratentorial primitive neuroectodermal tumor), T-cell lymphoma (such as cutaneous T-cell lymphoma), testicular cancer, throat cancer, thymoma (such as childhood thymoma), thymoma or thymic carcinoma, thyroid cancer (such as childhood thyroid cancer), trophoblastic tumor (such as gestational trophoblastic tumor), unknown primary site carcinoma (such as adult or childhood unknown primary site carcinoma), urethral cancer (such as endometrial uterine cancer), uterine sarcoma, vaginal cancer, visual pathway or hypothalamic glioma (such as childhood visual pathway or hypothalamic glioma), vulvar cancer, waldenström macroglobulinemia, or wilms tumor (such as childhood wilms tumor). In various embodiments, the cancer has metastasized or has not metastasized.

[0135] The cancer may or may not be a hormone related or dependent cancer (*e.g.,* an estrogen or androgen related cancer). Benign tumors or malignant tumors can be diagnosed, prognosed, stabilized, treated, or prevented using the methods of the present invention.

[0136] In some embodiments, the subject has a cancer syndrome. A cancer syndrome is a genetic disorder in which genetic mutations in one or more genes predispose the affected individuals to the development of cancers and may also cause the early onset of these cancers. Cancer syndromes often show not only a high lifetime risk of developing cancer, but also the development of multiple independent primary tumors. Many of these syndromes are caused by mutations in tumor suppressor genes, genes that are involved in protecting the cell from turning cancerous. Other genes that can be affected are DNA repair genes, oncogenes and genes involved in the production of blood vessels (angiogenesis). Common examples of inherited cancer syndromes are hereditary breast-ovarian cancer syndrome and hereditary non-polyposis colon cancer (Lynch syndrome).

[0137] A subject with one or more polymorphisms or mutations n K-ras, p53, BRA, EGFR, or HER2 may be administered a treatment that targets K-ras, p53, BRA, EGFR, or HER2, respectively.

[0138] The disclosed methods can be used to direct a therapeutic regimen. The polymorphism or mutation may be associated with altered response to a particular treatment (such as increased or decreased efficacy or side-effects). Therapies are available and under development that target specific mutations associated with various cancers, including lung cancer and ovarian cancer. It is known that therapeutics can be effective against targeted mutations such as CNVs. Example 5 herein, provides a Table of targeted therapeutics indicated by CNVs in particular genes, (see Table 20).

Analytical Methods

[0139] Methods for determining ploidy herein, typically include an analytical method that analyzes allelic data, such as allelic count sequencing data, regarding a plurality of SNPs, receives or generates imperfectly phased allelic information, and generates individual and joint probabilities for different ploidy states, to determine a ploidy state of a chromosomal region. Such analytical methods have been reported (See e.g. WO 2007/062164, WO 2012/108920, and WO 2015/164432) and can be used in methods provided herein. Surprisingly, presented herein is data that shows that by choosing SNPs that are found within haploblocks, increased performance of such SNP-based analytical methods, can be achieved.

[0140] In such analytical methods, individual probabilities can be generated using a set of models or hypothesis of both different ploidy states and average allelic imbalance fractions for the set of polymorphic loci. For example, in a particularly illustrative example, individual probabilities are generated by modeling ploidy states of a first homolog of the chromosome region and a second homolog of the chromosome region. The ploidy states that are modeled include the following:

1. (1) all cells have no deletion or amplification of the first homolog or the second homolog

of the chromosome region;

2. (2) at least some cells have a deletion of the first homolog or an amplification of the second homolog of the chromosome region; and
3. (3) at least some cells have a deletion of the second homolog or an amplification of the first homolog of the chromosome region.

[0141] It will be understood that the above models can also be referred to as hypothesis that are used to constrain a model. Therefore, demonstrated above are 3 hypothesis that can be used.

[0142] The average allelic imbalance fractions modeled can include any range of average allelic imbalance that includes the actual average allelic imbalance of the chromosomal region. For example, in certain illustrative embodiments, the range of average allelic imbalance that is modeled can be between 0, 0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.75, 1, 2, 2.5, 3, 4, and 5% on the low end, and 1, 2, 2.5, 3, 4, 5, 10, 15, 20, 25, 30, 40, 50, 60, 70 80 90, 95, and 99% on the high end. The intervals for the modeling with the range can be any interval depending on the computing power used and the time allowed for the analysis. For example, 0.01, 0.05, 0.02, or 0.1 intervals can be modeled.

[0143] In certain illustrative embodiments, the sample has an average allelic imbalance for the chromosomal region of between 0.4% and 5%. In certain embodiments, the average allelic imbalance is low. In these embodiments, average allelic imbalance is typically less than 10%. In certain illustrative embodiments, the allelic imbalance is between 0.25, 0.3, 0.4, 0.5, 0.6, 0.75, 1, 2, 2.5, 3, 4, and 5% on the low end, and 1, 2, 2.5, 3, 4, and 5% on the high end. In other exemplary embodiments, the average allelic imbalance is between 0.4, 0.45, 0.5, 0.6, 0.7, 0.8, 0.9, or 1.0 on the low end and 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.5, 2.0, 3.0, 4.0, or 5.0 on the high end. For example, the average allelic imbalance of the sample in an illustrative example is between 0.45 and 2.5%. In another example, the average allelic imbalance is detected with a sensitivity of 0.45, 0.5, 0.6, 0.8, 0.8, 0.9, or 1.0. In An exemplary sample with low allelic imbalance in methods of the present invention include plasma samples from individuals with cancer having circulating tumor DNA or plasma samples from pregnant females having circulating fetal DNA.

[0144] It will be understood that for SNVs, the proportion of abnormal DNA is typically measured using mutant allele frequency (number of mutant alleles at a locus / total number of alleles at that locus). Since the difference between the amounts of two homologs in tumours is analogous, we measure the proportion of abnormal DNA for a CNV by the average allelic imbalance (AAI), defined as |(H1 - H2)|/(H1 + H2), where Hi is the average number of copies of homolog i in the sample and Hi/(H1 + H2) is the fractional abundance, or homolog ratio, of homolog i. The maximum homolog ratio is the homolog ratio of the more abundant homolog.

[0145] Assay drop-out rate is the percentage of SNPs with no reads, estimated using all SNPs.

Single allele drop-out (ADO) rate is the percentage of SNPs with only one allele present, estimated using only heterozygous SNPs. Genotype confidence can be determined by fitting a binomial distribution to the number of reads at each SNP that were B-allele reads, and using the ploidy status of the focal region of the SNP to estimate the probability of each genotype.

[0146] Genotypic measurements are made during methods provided herein. Such measurements can be obtained by measuring signal intensities for different alleles for each of the SNPs using a SNP microarray or by allele frequency measurements using sequencing reactions, especially high throughput sequencing. Accordingly, genotypic measurements include allele frequency data and allele counts, for example. Genotypic measurements can be made by amplifying genetic material in the sample and then analyzing amplicons using SNP microarrays and/or high throughput sequencing.

[0147] In certain illustrative examples, the allele frequency data is corrected for errors before it is used to generate individual probabilities. In specific illustrative embodiments, the errors that are corrected include allele amplification efficiency bias. In other embodiments, the errors that are corrected include ambient contamination and genotyped contamination. In some embodiments, errors that are corrected include allele amplification bias, ambient contamination and genotype contamination. Analytical methods are provided herein, for correcting for such errors.

[0148] In certain embodiments, the individual probabilities are generated using a set of models of both different ploidy states and allelic imbalance fractions for the set of polymorphic loci. In these embodiments, and other embodiments, the joint probabilities are generated by considering the linkage between polymorphic loci on the chromosome region.

[0149] For tumor tissue samples, chromosomal aneuploidy (exemplified in this paragraph by CNVs) can be delineated by transitions between allele frequency distributions. In plasma samples, CNVs can be identified by a maximum likelihood algorithm that searches for plasma CNVs in regions where the tumor sample from the same individual also has CNVs, using haplotype information deduced from the tumor sample. This algorithm can model expected allelic frequencies across all allelic imbalance ratios at 0.025% intervals for three sets of hypotheses: (1) all cells are normal (no allelic imbalance), (2) some/all cells have a homolog 1 deletion or homolog 2 amplification, or (3) some/all cells have a homolog 2 deletion or homolog 1 amplification. The likelihood of each hypothesis can be determined at each SNP using a Bayesian classifier based on a beta binomial model of expected and observed allele frequencies at all heterozygous SNPs, and then the joint likelihood across multiple SNPs can be calculated, in certain illustrative embodiments taking linkage of the SNP loci into consideration, as exemplified herein. The maximum likelihood hypothesis can then be selected.

[0150] Consider a chromosomal region with an average of N copies in the tumor, and let c denote the fraction of DNA in plasma derived from the mixture of normal and tumor cells in a disomic region. AAI is calculated as:

$$AAI = \frac{c\,|\,N-2\,|}{2+c\left(N-2\right)}$$

[0151] In certain illustrative examples, the allele frequency data is corrected for errors before it is used to generate individual probabilities. Different types of error and/or bias correction are disclosed herein. In specific illustrative embodiments, the errors that are corrected are allele amplification efficiency bias. In other embodiments, the errors that are corrected include ambient contamination and genotype contamination. In some embodiments, errors that are corrected include allele amplification bias, ambient contamination and genotype contamination.

[0152] It will be understood that allele amplification efficiency bias can be determined for an allele as part of an experiment or laboratory determination that includes an on test sample, or it can be determined at a different time using a set of samples that include the allele whose efficiency is being calculated. Ambient contamination and genotype contamination are typically determined on the same run as the on-test sample analysis.

[0153] In certain embodiments, ambient contamination and genotype contamination are determined for homozygous alleles in the sample. It will be understood that for any given sample from an individual some loci in the sample, will be heterozygous and others will be homozygous, even if a locus is selected for analysis because it has a relatively high heterozygosity in the population. It is advantageous in some embodiments, although ploidy of a chromosomal region can be determined using heterozygous loci for an individual. Homozygous loci can be used to calculate ambient and genotype contamination.

[0154] In certain illustrative examples, the selecting is performed by analyzing a magnitude of a difference between the phased allelic information and estimated allelic frequencies generated for the models.

[0155] In illustrative examples, the individual probabilities of allele frequencies are generated based on a beta binomial model of expected and observed allele frequencies at the set of polymorphic loci. In illustrative examples, the individual probabilities are generated using a Bayesian classifier.

[0156] In certain illustrative embodiments, the nucleic acid sequence data is generated by performing high throughput DNA sequencing of a plurality of copies of a series of amplicons generated using a multiplex amplification reaction, wherein each amplicon of the series of amplicons spans at least one polymorphic loci of the set of polymorphic loci and wherein each of the polymeric loci of the set is amplified. In certain embodiments, the multiplex amplification reaction is performed under limiting primer conditions for at least ½ of the reactions. In some embodiments, limiting primer concentrations are used in 1/10, 1/5, ¼, 1/3, ½, or all of the reactions of the multiplex reaction. Provided herein are factors to consider to achieve limiting primer conditions in an amplification reaction such as PCR.

[0157] In certain embodiments, methods provided herein detect ploidy for multiple chromosomal regions across multiple chromosomes. Accordingly, the chromosomal ploidy in these embodiments is determined for a set of chromosome regions in the sample. For these embodiments, higher multiplex amplification reactions are needed. Accordingly, for these embodiments the multiplex amplification reaction can include, for example, between 2,500 and 50,000 multiplex reactions. In certain embodiments, the following ranges of multiplex reactions are performed: between 100, 200, 250, 500, 1000, 2500, 5000, 10,000, 20,000, 25000, 50000 on the low end of the range and between 200, 250, 500, 1000, 2500, 5000, 10,000, 20,000, 25000, 50000, and 100,000 on the high end of the range.

[0158] In illustrative embodiments, the set of polymorphic loci is a set of loci that are known to exhibit high heterozygosity. However, it is expected that for any given individual, some of those loci will be homozygous. In certain illustrative embodiments, methods of the invention utilize nucleic acid sequence information for both homozygous and heterozygous loci for an individual. The homozygous loci of an individual are used, for example, for error correction, whereas heterozygous loci are used for the determination of allelic imbalance of the sample. In certain embodiments, at least 10% of the polymorphic loci are heterozygous loci for the individual.

[0159] As disclosed herein, preference is given for analyzing target SNP loci that are known to be heterozygous in the population. Accordingly, in certain embodiments, polymorphic loci are chosen wherein at least 10, 20, 25, 50, 75, 80, 90, 95, 99, or 100% of the polymorphic loci are known to be heterozygous in the population.

[0160] In some examples, the method further comprises performing the method on a control sample with a known average allelic imbalance ratio. The control can have an average allelic imbalance ratio for a particular allelic state indicative of aneuploidy of the chromosome region, of between 0.4 and 10% to mimic an average allelic imbalance of an allele in a sample that is present in low concentrations, such as would be expected for a circulating free DNA from a fetus or from a tumor.

[0161] In certain embodiments of the methods of determining ploidy, the sample is a plasma sample from an individual suspected of having cancer. In these embodiments, the method further comprises determining based on the selecting whether copy number variation is present in cells of a tumor of the individual. For these embodiments, the sample can be a plasma sample from an individual. For these embodiments, the method can further include determining, based on the selecting, whether cancer is present in the individual.

[0162] These embodiments for determining ploidy of a chromosomal region, can further include detecting a single nucleotide variant at a single nucleotide variance location in a set of single nucleotide variance locations, wherein detecting either a chromosomal aneuploidy or the single nucleotide variant or both, indicates the presence of circulating tumor nucleic acids in the sample.

[0163] As disclosed herein, certain embodiments of the methods of determining ploidy can further include removing outliers from the initial or corrected allele frequency data before comparing the initial or the corrected allele frequencies to the set of models. For example, in certain embodiments, loci allele frequencies that are at least 2 or 3 standard deviations above or below the mean value for other loci on the chromosome region, are removed from the data before being used for the modeling.

[0164] As mentioned herein, it will be understood that for illustrative embodiments provided herein, including those for determining ploidy of a chromosomal region, imperfectly phased data is generated. It will also be understood, that provided herein are a number of features that provide improvements over prior methods for detecting ploidy, and that many different combinations of these features could be used. Furthermore, it will be understood that the plurality of polymorphic loci on a chromosome region can be linked loci since they are on the same chromosome region, and therefore have some statistical correlation for phasing estimates. However, within haploblocks, there is an increased statistical correlation of polymorphic loci with respect to phase estimation, because the loci exhibit a strong linkage disequilibrium, as disclosed herein.

[0165] In various embodiments, the phase of an individual's genetic data is estimated using data about the probability of chromosomes crossing over at different locations in a chromosome or chromosome region (such as using recombination data such as can be found in the HapMap database to create a recombination risk score for any interval) to model dependence between polymorphic alleles on the chromosome or chromosome region. In some embodiments, allele counts at the polymorphic loci are calculated on a computer based on sequencing data or SNP array data. A plurality of hypotheses each pertaining to a different possible state of the chromosome or chromosome region (such as an overrepresentation of the number of copies of a first homologous chromosome region as compared to a second homologous chromosome region in the genome of one or more cells from an individual, a duplication of the first homologous chromosome region, a deletion of the second homologous chromosome region, or an equal representation of the first and second homologous chromosome regions) can be created (such as creation on a computer); a model (such as a joint distribution model) for the expected allele counts at the polymorphic loci on the chromosome can be built (such as building on a computer) for each hypothesis; a relative probability of each of the hypotheses can be determined (such as determination on a computer) using the j oint distribution model and the allele counts; and the hypothesis with the greatest probability can be selected. In some embodiments, building a joint distribution model for allele counts and the step of determining the relative probability of each hypothesis are done using a method that does not require the use of a reference chromosome.

[0166] In some embodiments, the analytical methods utilize a statistical 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. In some embodiments, the analytical methods 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. In some embodiments, the ratio of target DNA to total DNA in the sample utilizes maximum likelihood estimation, maximum a-posteriori estimation, Bayesian estimation, dynamic estimation (such as dynamic Bayesian estimation), and/or expectation-maximization estimation.

[0167] 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 region as compared to a second homologous chromosome region 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 region or the deletion of the second homologous chromosome region. In some embodiments, there is not an overrepresentation since the first and homologous chromosome regions 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. In this specification the phrase "a first homologous chromosome region as compared to a second homologous chromosome region" means a first homolog of a chromosome region and a second homolog of the chromosome region.

[0168] In some embodiments, the method further 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 region) divided by the measured quantity of one or more other alleles (such as the allele on the second homologous chromosome region) for the locus. The calculated allele ratios can 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).

[0169] In one embodiment, an individual's genetic data is phased using an algorithm that estimates haplotypes from genotype data, such as an algorithm that uses localized haplotype clustering (*see, e.g.,* Browning and Browning, "Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering" Am J Hum Genet. Nov 2007; 81(5): 1084-1097,. An exemplary program is Beagle version: 3.3.2 or version 4 (available at the world wide web at hfaculty.washington.edu/browning/beagle/beagle.html,

[0170] In one embodiment, an individual's genetic data is phased using an algorithm that estimates haplotypes from genotype data, such as an algorithm that uses the decay of linkage disequilibrium with distance, the order and spacing of genotyped markers, missing-data imputation, recombination rate estimates, or a combination thereof (*see, e.g.,* Stephens and Scheet, "Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing-Data Imputation" Am. J. Hum. Genet. 76:449-462, 2005,. An exemplary program is PHASE v.2.1 or v2.1.1. (available at the world wide web at stephenslab.uchicago.edu/software.html,.

[0171] In one embodiment, an individual's genetic data is phased using an algorithm that estimates haplotypes from population genotype data, such as an algorithm that allows cluster memberships to change continuously along the chromosome according to a hidden Markov model. This approach is flexible, allowing for both "block-like" patterns of linkage disequilibrium and gradual decline in linkage disequilibrium with distance *(see, e.g.,* Scheet and Stephens, "A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase." Am J Hum Genet, 78:629-644, 2006,). An exemplary program is fastPHASE (available at the world wide web at stephenslab.uchicago.edu/software.html,.

[0172] In one embodiment, an individual's genetic data is phased using a genotype imputation method, such as a method that uses one or more of the following reference datasets: HapMap dataset, datasets of controls genotyped on multiple SNP chips, and densely typed samples from the 1,000 Genomes Project. An exemplary approach is a flexible modelling framework that increases accuracy and combines information across multiple reference panels (*see, e.g.,* Howie, Donnelly, and Marchini (2009) "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." PLoS Genetics 5(6): e1000529, 2009,. Exemplary programs are IMPUTE or IMPUTE version 2 (also known as IMPUTE2) (available at the world wide web at mathgen.stats.ox.ac.uk/impute/impute_v2.html,.

[0173] In one embodiment, an individual's genetic data is phased using an algorithm that infers haplotypes, such as an algorithm that infers haplotypes under the genetic model of coalescence with recombination, such as that developed by Stephens in PHASE v2.1. The major algorithmic improvements rely on the use of binary trees to represent the sets of candidate haplotypes for each individual. These binary tree representations: (1) speed up the computations of posterior probabilities of the haplotypes by avoiding the redundant operations made in PHASE v2.1, and (2) overcome the exponential aspect of the haplotypes inference problem by the smart exploration of the most plausible pathways *(i.e.,* haplotypes) in the binary trees (*see, e.g.,* Delaneau, Coulonges and Zagury, "Shape-IT: new rapid and accurate algorithm for haplotype inference," BMC Bioinformatics 9:540, 2008 doi:10.1186/1471-2105-9-540,. An exemplary program is SHAPEIT (available at the world wide web at mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html,.

[0174] In one embodiment, an individual's genetic data is phased using an algorithm that estimates haplotypes from population genotype data, such as an algorithm that uses haplotype-fragment frequencies to obtain empirically based probabilities for longer haplotypes. In some embodiments, the algorithm reconstructs haplotypes so that they have maximal local coherence (*see, e.g.,* Eronen, Geerts, and Toivonen, "HaploRec: Efficient and accurate large-scale reconstruction of haplotypes," BMC Bioinformatics 7:542, 2006,. An exemplary program is HaploRec, such as HaploRec version 2.3. (available at the world wide web at cs.helsinki.fi/group/genetics/haplotyping.html,.

[0175] In one embodiment, an individual's genetic data is phased using an algorithm that estimates haplotypes from population genotype data, such as an algorithm that uses a partition-ligation strategy and an expectation-maximization-based algorithm (*see, e.g.,* Qin, Niu, and Liu, "Partition-Ligation-Expectation-Maximization Algorithm for Haplotype Inference with Single-Nucleotide Polymorphisms," Am J Hum Genet. 71(5): 1242-1247, 2002,. An exemplary program is PL-EM (available at the world wide web at people.fas.harvard.edu/ junliu/plem/click.html,.

[0176] In one embodiment, an individual's genetic data is phased using an algorithm that estimates haplotypes from population genotype data, such as an algorithm for simultaneously phasing genotypes into haplotypes and block partitioning. In some embodiments, an expectation-maximization algorithm is used (*see, e.g.,* Kimmel and Shamir, "GERBIL: Genotype Resolution and Block Identification Using Likelihood," Proceedings of the National Academy of Sciences of the United States of America (PNAS) 102: 158-162, 2005,. An exemplary program is GERBIL, which is available as part of the GEVALT version 2 program (available at the world wide web at acgt.cs.tau.ac.il/gevalt/,.

[0177] In one embodiment, an individual's genetic data is phased using an algorithm that estimates haplotypes from population genotype data, such as an algorithm that uses an EM algorithm to calculate ML estimates of haplotype frequencies given genotype measurements which do not specify phase. The algorithm also allows for some genotype measurements to be missing (due, for example, to PCR failure). It also allows multiple imputation of individual haplotypes (*see, e.g.,* Clayton, D. (2002), "SNPHAP: A Program for Estimating Frequencies of Large Haplotypes of SNPs",. An exemplary program is SNPHAP (available at the world wide web at gene.cimr.cam.ac.uk/clayton/software/snphap.txt,.

[0178] In one embodiment, an individual's genetic data is phased using an algorithm that estimates haplotypes from population genotype data, such as an algorithm for haplotype inference based on genotype statistics collected for pairs of SNPs. This software can be used for comparatively accurate phasing of large number of long genome sequences, e.g. obtained from DNA arrays. An exemplary program takes genotype matrix as an input, and outputs the corresponding haplotype matrix (*see, e.g.,* Brinza and Zelikovsky, "2SNP: scalable phasing based on 2-SNP haplotypes," Bioinformatics.22(3):371-3, 2006,. An exemplary program is 2SNP (available at the world wide web at alla.cs.gsu.edu/~software/2SNP,.

[0179] Accordingly, in certain embodiments, publicly available programs, such as those disclosed above, can be utilized to estimate the phase genetic data such as allele frequency data from the sample. The Examples provided herein utilize imperfect haplotyping and demonstrate that haplotyping is more accurate within haploblocks. Therefore, by choosing loci within haploblocks for analysis of ploidy (e.g. CNV detection, ploidy determination, or AAI determination), improved results are obtained from those using imperfectly phased information that is from outside haploblocks. These methods for estimating phase of genetic data provided by the various methods disclosed herein, when used in illustrative embodiments, provide the value for c that is used in the Combined_Likelihoods equation provided herein.

[0180] In some embodiments, outside the scope of the claims, the method involves determining if there is an overrepresentation of the number of copies of the first homologous chromosome region 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 regions are present in equal proportions. The expected allele ratio assumes the possible alleles for a locus have an equal likelihood of being present. When 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. The expected allele ratio may be the same for all the loci, such as 0.5 for all loci. The expected allele ratio may assume 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/; 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,. The expected allele ratio may be 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 region. For example, the expected allele ratio for a particular individual can 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. With respect to prenatal testing, the expected allele ratio may be 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 region. For example, the expected allele ratio for the mixed sample can 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). The expected allele ratios can 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) or methods provided in U.S. Publication No 2012/0270212, filed Nov. 18, 2011.

[0181] A calculated allele ratio is indicative of an overrepresentation of the number of copies of the first homologous chromosome region 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. A calculated allele ratio is only considered indicative of overrepresentation if it is significantly greater or lower than the expected ratio for that locus. A calculated allele ratio is indicative of no overrepresentation of the number of copies of the first homologous chromosome region 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. Calculated ratios equal to the corresponding expected ratio may be ignored (since they are indicative of no overrepresentation).

[0182] 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.

[0183] In some embodiments of the above, 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 region is due to a duplication of the first homologous chromosome region or a deletion of the second homologous chromosome region in the genome of one or more of the cells.

[0184] An overrepresentation of the number of copies of the first homologous chromosome region 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 region 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 region 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 region 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 region 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 region 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 region is above a threshold value (indicative of a poor data fit).

[0185] An overrepresentation of the number of copies of the first homologous chromosome region 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 region 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 region 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 region 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 region 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 region 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 region 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.

[0186] Determining if there is an overrepresentation of the number of copies of the first homologous chromosome region includes enumerating a set of one or more hypotheses specifying the degree of overrepresentation of the first homologous chromosome region. On exemplary hypothesis is the absence of an overrepresentation since the first and homologous chromosome regions are present in equal proportions (such as one copy of each segment in a diploid sample). Other exemplary hypotheses include the first homologous chromosome region 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 region). Another exemplary hypothesis includes the deletion of the second homologous chromosome region. Yet another exemplary hypothesis is the deletion of both the first and the second homologous chromosome regions. 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.

Exemplary Methods for Predicting Allele Ratios

[0187] Exemplary methods are discussed below for calculating expected allele ratios for a sample. Table 3 shows expected allele ratios for a mixed sample (such as a maternal blood sample) containing nucleic acids from both the mother and the fetus. These expected allele ratios indicate what is expected for measurement of the total amount of each allele, including the amount of the allele from both maternal nucleic acids and fetal nucleic acids in the mixed sample. In an example, the mother is heterozygous at two neighboring loci that are expected to cosegregate (e.g., two loci for which no chromosome crossovers are expected between the loci). Thus, the mother is (AB, AB). Now imagine that the phased data for the mother indicates that for one haplotype she is (A, A); thus, for the other haplotype one can infer that she is (B, B). Table 3 gives the expected allele ratios for different hypotheses where the fetal fraction is 20%. For this example, no knowledge of the paternal data is assumed, and the heterozygosity rate is assumed to be 50%. The expected allele ratios are given in terms of (expected proportion of A reads / total number of reads) for each of the two SNPs. These ratios are calculated both using maternal phased data (the knowledge that one haplotype is (A, A) and one is (B, B)) and without using the maternal phased data. Table 3 includes different hypotheses for the number of copies of the chromosome region in the fetus from each parent.

Table 3: Expected Genetic Data for Mixed Sample of Maternal and Fetal Nucleic Acids

Copy Number Hypothesis	**Expected allele ratios when using maternal phased data**	**Expected allele ratios when not using maternal phased data**		
Monosomy (maternal copy missing)	(0.444; 0.444)	(0.444; 0.444)		
	(0.444; 0.555)	(0.444; 0.555)		
	(0.555; 0.444)	(0.555; 0.444)		
	(0.555; 0.555)	(0.555; 0.555)		
Monosomy (paternal copy missing)	(0.444; 0.444)	(0.444; 0.444)		
	(0.555; 0.555)	(0.444; 0.555)		
		(0.555; 0.444)		
		(0.555; 0.555)		
Disomy	(0.40; 0.40)	(0.40; 0.40)	(0.50; 0.60)	
	(0.40; 0.50)	(0.40; 0.50)	(0.60; 0.40)	
	(0.50; 0.40)	(0.40; 0.60)	(0.60; 0.50)	
	(0.50; 0.50)	(0.50; 0.40)	(0.60; 0.60)	
	(0.50; 0.60)	(0.50; 0.50)		
	(0.60; 0.50) (0.60; 0.60)			
Trisomy (extra matching maternal copy)	(0.36; 0.36)		(0.36; 0.36)	(0.54; 0.36)
	(0.36; 0.45)		(0.36; 0.45)	(0.54; 0.45)
	(0.45; 0.36)		(0.36; 0.54)	(0.54; 0.54)
	(0.45; 0.45)		(0.36; 0.63)	(0.54; 0.63)

Copy Number Hypothesis	Expected allele ratios when using maternal phased data	Expected allele ratios when not using maternal phased data		
	(0.54; 0.54)		(0.45; 0.36)	(0.63; 0.36)
	(0.54; 0.63)		(0.45; 0.45)	(0.63; 0.45)
	(0.63; 0.54)		(0.45; 0.54)	(0.63; 0.54)
	(0.63; 0.63)		(0.45; 0.63)	(0.63; 0.63)
Trisomy (extra unmatching maternal copy)	(0.45, 0.45)		(0.36; 0.36)	(0.54; 0.36)
	(0.45; 0.54)		(0.36; 0.45)	(0.54; 0.45)
	(0.54; 0.45)		(0.36; 0.54)	(0.54; 0.54)
	(0.54; 0.54)		(0.36; 0.63)	(0.54; 0.63)
			(0.45; 0.36)	(0.63; 0.36)
			(0.45; 0.45)	(0.63; 0.45)
			(0.45; 0.54)	(0.63; 0.54)
			(0.45; 0.63)	(0.63; 0.63)
Trisomy (extra matching paternal copy)	(0.36; 0.36)		(0.36; 0.36)	(0.54; 0.36)
	(0.36; 0.54)		(0.36; 0.45)	(0.54; 0.45)
	(0.54; 0.36)		(0.36; 0.54)	(0.54; 0.54)
	(0.54; 0.54)		(0.36; 0.63)	(0.54; 0.63)
	(0.45; 0.45)		(0.45; 0.36)	(0.63; 0.36)
	(0.45; 0.63)		(0.45; 0.45)	(0.63; 0.45)
	(0.63; 0.45)		(0.45; 0.54)	(0.63; 0.54)
	(0.63; 0.63)		(0.45;	(0.63;

Copy Number Hypothesis	Expected allele ratios when using maternal phased data	Expected allele ratios when not using maternal phased data		
			0.63)	0.63)
Trisomy (extra unmatching paternal copy)	(0.36; 0.36)	(0.54; 0.36)	(0.36; 0.36)	(0.54; 0.36)
	(0.36; 0.45)	(0.54; 0.45)	(0.36; 0.45)	(0.54; 0.45)
	(0.36; 0.54)	(0.54; 0.54)	(0.36; 0.54)	(0.54; 0.54)
	(0.36; 0.63)	(0.54; 0.63)	(0.36; 0.63)	(0.54; 0.63)
	(0.45; 0.36)	(0.63; 0.36)	(0.45; 0.36)	(0.63; 0.36)
	(0.45; 0.45)	(0.63; 0.45)	(0.45; 0.45)	(0.63; 0.45)
	(0.45; 0.54)	(0.63; 0.54)	(0.45; 0.54)	(0.63; 0.54)
	(0.45; 0.63)	(0.63; 0.63)	(0.45; 0.63)	(0.63; 0.63)

[0188] In addition to the fact that using phased data reduces the number of possible expected allele ratios, it also changes the prior likelihood of each of the expected allele ratios, such that the maximum likelihood result is more likely to be correct. Eliminating expected allele ratios or hypotheses that are not possible increases the likelihood that the correct hypothesis will be chosen. As an example, suppose the measured allele ratios are (0.41, 0.59). Without using phased data, one might assume that the hypothesis with maximum likelihood is a disomy hypothesis (given the similarity of the measured allele ratios to expected allele ratios of (0.40, 0.60) for disomy). However, using phased data, one can exclude (0.40, 0.60) as expected allele ratios for the disomy hypothesis, and one can select a trisomy hypothesis as more likely.

[0189] Assume the measured allele ratios are (0.4, 0.4). Without any haplotype information, the probability of a maternal deletion at each SNP would be 0.5 x P(A deleted) + 0.5 x P(B deleted). Therefore, although it looks like A is deleted (missing in the fetus), the likelihood of deletion would be the average of the two. For high enough fetal fraction, one can still determine the most likely hypothesis. For low enough fetal fraction, averaging may work in disfavor of the deletion hypothesis. However, with haplotype information, the probability of homolog 1 being deleted, P(A deleted), is greater and will fit the measured data better. If desired, crossover probabilities between the two loci can also be considered.

Further Detailed Exemplary Embodiments of Analytical Methods

Exemplary Test Statistic for Analysis of Phased Data

[0190] 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. Letfdenote 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. In other embodiments,fdenotes the fraction of ctDNA DNA, RNA, or cells in a mixture of ctDNA and DNA, RNA, or cells from non-cancerous cells of the individual. 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.

[0191] 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 AB or BA are analyzed. Let N_i denote the number of sequence reads of the ith SNP, and A_i and B_i denote the number of reads of the ith SNP that indicate allele A and B, respectively. It is assumed:

$$N_i = A_i + B_i.$$

[0192] The allele ratio R_i is defined:

$$R_i \triangleq \frac{A_i}{N_i}.$$

[0193] Let T denote the number of SNPs targeted.

[0194] Without loss of generality, some embodiments focus on a single chromosome region. As a matter of further clarity, in this specification the phrase "a first homologous chromosome region as compared to a second homologous chromosome region" means a first homolog of a chromosome region and a second homolog of the chromosome region. In some such embodiments, all of the target SNPs are contained in the segment chromosome of interest. In other embodiments, multiple chromosome regions are analyzed for possible copy number variations.

Map Estimation

[0195] 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

$$X_i \triangleq \begin{cases} 1 & R_i < 0.5\ and\ SNP\ i\ AB \\ 0 & R_i \geq 0.5\ and\ SNP\ i\ AB \\ 0 & R_i < 0.5\ and\ SNP\ i\ BA \\ 1 & R_i \geq 0.5\ and\ SNP\ i\ BA \end{cases}$$

[0196] Then define

$$S \triangleq \sum_{All\ SNPs} X_i.$$

[0197] The distributions of the X_i 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

[0198] Under the hypothesis that the target segment is not deleted or duplicated,

$$X_i = \begin{cases} 0 & wp\ 1 - p\left(\frac{1}{2}, N_i\right) \\ 1 & wp\ p\left(\frac{1}{2}, N_i\right) \end{cases}$$

where

$$p(b,n) \triangleq Pr\left\{X \sim Bino(b,n) \geq \frac{n}{2}\right\}.$$

[0199] If we assume a constant depth of read N, this gives us a Binomial distribution S with parameters

$$p\left(\frac{1}{2}, N\right)$$

and T.

Deletion Hypotheses

[0200] Under the hypothesis that the first homolog is deleted *(i.e.,* an AB SNP becomes B, and a BA SNP becomes A), then R_i has a Binomial distribution with parameters

$$1 - \frac{1}{2-f}$$

and T for AB SNPs, and

$$\frac{1}{2-f}$$

and T for BA SNPs. Therefore,

$$X_i = \begin{cases} 0 & wp\ 1 - p\left(\frac{1}{2-f}, N_i\right) \end{cases}$$

$$\begin{cases} 1 & wp\, p\left(\frac{1}{2-f}, N_i\right) \end{cases}$$

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

$p\left(\frac{1}{2-f}, N\right)$

and T.

[0201] Under the hypothesis that the second homolog is deleted *(i.e.,* an AB SNP becomes A, and a BA SNP becomes B), then *R_i* has a Binomial distribution with parameters

$\frac{1}{2-f}$

and T for AB SNPs, and

$1-\frac{1}{2-f}$

and T for BA SNPs. Therefore,

$$X_i = \begin{cases} 0 & wp\, p\left(\frac{1}{2-f}, N_i\right) \\ 1 & wp\ 1-p\left(\frac{1}{2-f}, N_i\right) \end{cases}$$

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

$1-p\left(\frac{1}{2-f}, N\right)$

and T.

Duplication Hypotheses

[0202] Under the hypothesis that the first homolog is duplicated *(i.e.,* an AB SNP becomes AAB, and a BA SNP becomes BBA), then R_i has a Binomial distribution with parameters

$\frac{1+f}{2+f}$

and T for AB SNPs, and

$1-\frac{1+f}{2+f}$

and T for BA SNPs. Therefore,

$$X_i = \begin{cases} 0 & wp\, p\left(\frac{1+f}{2+f}, N_i\right) \\ 1 & wp\ 1-p\left(\frac{1+f}{2+f}, N_i\right) \end{cases}$$

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

$1-p\left(\frac{1+f}{2+f}, N\right)$

and T.

[0203] Under the hypothesis that the second homolog is duplicated *(i.e.,* an AB SNP becomes ABB, and a BA SNP becomes BAA), then R_i has a Binomial distribution with parameters

$1-\frac{1+f}{2+f}$

and T for AB SNPs, and

$\frac{1+f}{2+f}$

and T for BA SNPs. Therefore,

$$X_i = \begin{cases} 0 & wp\ 1-p\left(\dfrac{1+f}{2+f}, N_i\right) \\ 1 & wp\ p\left(\dfrac{1+f}{2+f}, N_i\right) \end{cases}$$

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

$p\left(\frac{1+f}{2+f}, N\right)$

and T.

Classification

[0204] As demonstrated in the sections above, X_i is a binary random variable with

$$Pr\{X_1 = 1\} = \begin{cases} p\left(\frac{1}{2}, N_i\right) & given\ disomy \\ p\left(\frac{1}{2-f}, N_i\right) & homolog\ 1\ deletion \\ 1-p\left(\frac{1}{2-f}, N_i\right) & homolog\ 2\ deletion \\ 1-p\left(\frac{1+f}{2+f}, N_i\right) & homolog\ 1\ duplication \\ p\left(\frac{1+f}{2+f}, N_i\right) & homolog\ 2\ duplication \end{cases}$$

[0205] 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

$$S \sim \begin{cases} Bino\left(p\left(\dfrac{1}{2}, N\right), T\right) & given\ disomy \\ Bino\left(p\left(\dfrac{1}{2-f}, N\right), T\right) & homolog\ 1\ deletion \\ Bino\left(1-p\left(\dfrac{1}{2-f}, N\right), T\right) & homolog\ 2\ deletion \\ Bino\left(1-p\left(\dfrac{1+f}{2+f}, N\right), T\right) & homolog\ 1\ duplication \\ Bino\left(p\left(\dfrac{1+f}{2+f}, N\right), T\right) & homolog\ 2\ duplication \end{cases}$$

[0206] 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 regions 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.

[0207] 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 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.

[0208] 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,. 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,. 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, that assumes certain loci are methylated or preferentially methylated in the fetus, and those same loci are unmethylated or preferentially unmethylated in the mother.

[0209] 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

[0210] 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. The probability of *S* under the disomy hypothesis may be 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). This indicates that a duplication or deletion of the chromosome region is present. If desired, the false positive rate can be altered by adjusting the threshold value.

Illustrative Methods for Analysis of Phased Data

[0211] 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. 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. 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.

[0212] The following paragraphs set out a specific non-limiting example of specific analytical considerations for practicing a quantitative, allelic method of the present invention, for determining copy number, ploidy, AAI and/or detecting aneuploidy and/or CNV, referred to herein as the Allelic _Analysis_Example. The following hypotheses are considered: H_{11} (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_{21} (presence of cells with homolog 1 duplication), H_{12} (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:

Equation (1):

$$
\begin{aligned}
r(AB,H_{11}) = r(BA,H_{11}) &= 0.5, \\
r(AB,H_{10}) = r(BA,H_{01}) &= \frac{1}{2-f}, \\
& \quad 1-f
\end{aligned}
$$

$$\begin{aligned} r(AB,H_{01}) = r(BA,H_{10}) &= \frac{\text{[illegible]}}{2-f}, \\ r(AB,H_{21}) = r(BA,H_{12}) &= \frac{1+f}{2+f}, \\ r(AB,H_{12}) = r(BA,H_{21}) &= \frac{1}{2+f}. \end{aligned}$$

Bias, Contamination, and Sequencing Error Correction:

[0213] A method of the invention, such as the exemplary Allelic _Analysis__Example, can then consider bias, contamination and sequencing error correction. For example, the observation that D_S at the SNP can include the number of original mapped reads with each allele present, $n_A{}^0$ and $n_B{}^0$. Then, one can find the corrected reads n_A and n_B using the expected bias in the amplification of *A* and *B* alleles.

[0214] Let c_a denote the ambient contamination (such as contamination from DNA in the air or environment) and $r(c_a)$ to denote the allele ratio for the ambient contaminant (which is taken to be 0.5 initially). Moreover, c_g denotes the genotyped contamination rate (such as the contamination from another sample), and $r(c_g)$ is the allele ratio for the contaminant. Let $s_e(A,B)$ and $s_e(B,A)$ 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).

[0215] One can find the observed allele ratio $q(r,\ c_a,\ r(c_a)\ ,\ c_g,\ r(c_g),\ s_e(A,B),\ s_e(B,A)\ )$ for a given expected allele ratio *r* by correcting for ambient contamination, genotyped contamination, and sequencing error.

[0216] 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 can be referred to as a reference allele). Then, we have $P(r(c_g) = 0) = (1\text{-}p)^2$, $P(r(cg) = 0) = 2p(1\text{-}p)$, *and* $P(r(cg) = 0) =p^2$. The conditional expectation over $r(c_g)$ can be used to determine the $\boldsymbol{E}[q(r,\ c_a,\ r(c_a)\ ,\ c_g\ ,\ r(c_g),\ s_e(A,B),\ s_e(B,A))\ ]$. 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:

[0217] In the methods provided herein, a likelihood at each SNP can be determined. The equation below, Equation (2), gives the probability using a binomial analysis of observing n_A

and n_B given an allele ratio *r*:

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

[0218] Let D_s denote the data for SNP *s*. For each hypothesis h ∈ { H_{11}, H_{01}, H_{10}, H_{21}, H_{12} }, one can let *r=r(AB,h)* or *r=r(BA,h)* in the equation (1) and find the conditional expectation over *r(cg)* to determine the observed allele ratio $E(q(r, c_a, r(c_a), c_g, r(c_g))$]. Then, letting $r = E[q(r, c_a, r(c_a), c_g, r(c_g), s_e(A,B), s_e(B,A))$] in equation (2) one can determine $P(D_s|h,f)$.

[0219] Methods of the present invention, such as the Allelic _Analysis__Example, can use a beta-binomial distribution. Equation (3) gives the likelihood of observing n_A and n_B given an expected allele ratio r following a beta distribution with parameters α and β. α and β are estimated from the training data.

$$Lik(n_A,n_B|r) = \binom{n_A+n_B}{n_A} \frac{\Gamma(n_A+\alpha)\Gamma(n_B+\beta)}{\Gamma(n_A+n_B+\alpha+\beta)} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \quad (3)$$

Search Algorithm:

[0220] Methods of the present invention can then use a search algorithm to search for the average allelic imbalance value that has the highest likelihood of being correct. In some examples of methods provided herein, such as the Allelic _Analysis__Example, SNPs with allele ratios that seem to be outliers can be ignored (such as by ignoring or eliminating SNPs with allele ratios that are at least 2 or 3 standard deviations above or below the mean value). Note that an advantage identified for this approach is that in the presence of higher mosaicism percentage, the variability in the allele ratios can be high, hence this ensures that SNPs will not be trimmed due to mosaicism.

[0221] In methods of the present invention, such as such as the Allelic _Analysis_Example method, $F = \{f_1, ..., f_N\}$ can denote the search space for the mosaicism percentage (such as the tumor fraction). The method can determine $P(D_s|h,f)$ at each SNP *s* and *f* ε *F*, and combine the likelihood over all SNPs.

[0222] 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* ∈ *F** can be determined. If desired, the confidence for each hypothesis can be determined.

Combining Likelihoods

[0223] Methods provided herein can combine likelihoods using phased date. For example, in the Allelic _Analysis_Example method, likelihoods using phased data, consider two consecutive SNPs s1 and s2, and use D1 and D2 to denote the allele data in these SNPs. Provided herein is an example on how, as incorporated into the Allelic _Analysis_Example, to combine the likelihoods for these two SNPs. Let c denote the probability that two consecutive heterozygous SNPs have the same allele in the same homolog (i.e., both SNPs are AB or both SNPs are BA). Hence 1-c is the probability that one SNP is AB and the other one is BA. For example, consider the hypothesis H10 and allelic imbalance value f. First, assume that all likelihoods are computed assuming that all SNPs are either AB or BA. Then, we can combine the likelihoods in two consecutive SNPs in the following fomula (Combined_Likelihoods):

$$Lik(D_1, D_2 \,|H_{\mathbf{10}}, f) =$$

$$Lik(D_1 \,|H_{10}, f) \times c \times Lik(D_2 \,|H_{10}, f) + Lik(D_1 \,|H_{10}, f) \times (1 - c) \times Lik(D_2 \,|H_{01}, f).$$

The above can be done recursively to determine the joint likelihood $Lik(D_1, \ldots, D_N \,|H_{10}, f)$ for all SNPs.

[0224] It is noteworthy that c values can be obtained as outputs from informatics haplotyping programs, as disclosed herein. In the presence of perfect haplotype information, we have c=0 or c=1 for individual haploblocks. In the absence of perfect haplotype information, but where target polymorphic loci are selected within haploblocks, estimates of haplotyping are improved, and therefore c values are closer to 0 or 1 than when polymorphic loci are analyzed that are not within haploblocks. Therefore, it is believed, and demonstrated in the Examples herein both in computer simulation and actual wet lab data, that by choosing loci within haploblocks, combined likelihoods can yield sufficiently accurate estimates of average allelic imbalance, chromosome copy number, and CNV, even with using estimated phase information that is not perfect. This accuracy is improved as more polymorphic loci within a chromosome region of interest are analyzed. This improved accuracy of determining and/or detecting average allelic imbalance, chromosome copy number, and/or CNV is especially useful in embodiments where average allelic imbalance in a sample is between 1, 2, or 3% on the low end of the range and 40, 30, 25 or 20% on high end of the range.

Theoretical Perfomance using Simulations:

[0225] If desired, one can evaluate the theoretical performance of a method provided herein 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.

First Simulation Experiment:

[0226] Accordingly, we evaluated the theoretical performance of the Allelic _Analysis_Example method. The experiment focused on $S \in \{500, 1000\}$, $D \varepsilon \{500, 1000\}$ and $p \in \{0\%, 1\%, 2\%, 3\%, 4\%, 5\%\}$. We performed 1,000 simulation experiments in each 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 *ofp>0%)* were determined both with or without phase information. Including phase information was very helpful in reducing false positive rates, 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.

[0227] 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_{10} and $H_{01},$ 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:

[0228] We then evaluated the theoretical performance of the Allelic Analysis Example method in a second simulation. This experiment focused on $S \in \{100, 200, 300, 400, 500\}$, $D \in \{1000, 2000, 3000, 4000, 5000\}$ *and* $p \in \{0\%, 1\%,1.5\%, 2\%,2.5\%, 3\%\}$ and 10000 random experiments at each setting. 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. The false negative rate is below 10% for $D \geq 3000$ and $N \geq 200$ using haplotype information, whereas the same performance is reached for *D=5000* and *N≥400.* The difference between the false negative rate was particularly stark for small mosaicism percentages. For example, when *p=1%,* a less than 20% false negative rate is never reached without haplotype data, whereas it is close to 0% for $N \geq 300$ and $D \geq 3000.$ For *p=3%,* a 0% false negative rate is observed with haplotype data, while $N \geq 300$ and $D \geq 3000$ is needed to reach the same performance without haplotype data.

Additional analytical method considerations:

[0229] In some embodiments, a beta binomial distribution is used instead of binomial distribution. In some embodiments, a reference chromosome or chromosome region is used to determine the sample specific parameters of beta binomial.

Exemplary Reference Chromosome Segments

[0230] In some embodiments, the one or more loci used to determine the tumor fraction are on a reference chromosomes segment, such as a chromosome region known or expected to be disomic, a chromosome region 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 region 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 region is disomic in both the cancer cells and noncancerous cells. In some embodiments, one or more chromosomes segments for which the confidence for a disomy call is high are used.

[0231] 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%).

[0232] 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.

[0233] 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.

[0234] 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.

[0235] 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.

[0236] 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.

[0237] 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.

[0238] 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.

[0239] In some embodiments, the value of f (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 off (see, e.g., US Publication No 2012/0190020; US Publication No 2012/0190021; US Publication No 2012/0190557; US Publication No 2012/0191358,. 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 f 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 f for various possible CNVs (such as one or two extra copies of a chromosome segment of interest).

[0240] 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 ½ x 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) x 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

[0241] 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.

[0242] In some embodiments, a subset of the samples with a sufficiently low likelihood of having cancer can 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 can 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 can 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 can 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, can 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 can be analyzed using a method for the purpose of detecting the presence of a CNV at one or more chromosomal regions.

[0243] 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.

[0244] 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 overrepresentation 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.

[0245] In some embodiments, allelic data is obtained, wherein the allelic data includes quantitative measurement(s) indicative of the number of copies of a specific allele of a polymorphic locus. In some embodiments, the allelic data includes quantitative measurement(s) indicative of the number of copies of each of the alleles observed at a polymorphic locus. Typically, quantitative measurements are obtained for all possible alleles of the polymorphic locus of interest. For example, any of the methods discussed in the preceding paragraphs for determining the allele for a SNP locus, such as for example, microarrays, qPCR, DNA sequencing, such as high throughput DNA sequencing, can be used to generate quantitative measurements of the number of copies of a specific allele of a polymorphic locus. This quantitative measurement is referred to herein as allelic frequency data or measured genetic allelic data. Methods using allelic data are sometimes referred to as quantitative allelic methods; this is in contrast to quantitative methods which exclusively use quantitative data from non-polymorphic loci, or from polymorphic loci but without regard to allelic identity. When the allelic data is measured using high-throughput sequencing, the allelic data typically include the number of reads of each allele mapping to the locus of interest.

[0246] In some embodiments, non-allelic data is obtained, wherein the non-allelic data includes quantitative measurement(s) indicative of the number of copies of a specific locus. The locus can be polymorphic or non-polymorphic. In some embodiments when the locus is non-polymorphic, the non-allelic data does not contain information about the relative or absolute quantity of the individual alleles that can be present at that locus. Methods using non-allelic data only (that is, quantitative data from non-polymorphic alleles, or quantitative data from polymorphic loci but without regard to the allelic identity of each fragment) are referred to as quantitative methods. Typically, quantitative measurements are obtained for all possible alleles of the polymorphic locus of interest, with one value associated with the measured quantity for all of the alleles at that locus, in total. Non-allelic data for a polymorphic locus can be obtained by summing the quantitative allelic for each allele at that locus. When the allelic data is measured using high-throughput sequencing, the non-allelic data typically includes the number of reads of mapping to the locus of interest. The sequencing measurements could indicate the relative and/or absolute number of each of the alleles present at the locus, and the non-allelic data includes the sum of the reads, regardless of the allelic identity, mapping to the locus. In some embodiments the same set of sequencing measurements can be used to yield both allelic data and non-allelic data. In some embodiments, the allelic data is used as part of a method to determine copy number at a chromosome of interest, and the produced non-allelic data can be used as part of a different method to determine copy number at a chromosome of interest. In some embodiments, the two methods are statistically orthogonal, and are combined to give a more accurate determination of the copy number at the chromosome of interest.

[0247] In any of the embodiments provided herein, methods of the invention can include a quantitative method for determining copy number or ploidy, or detecting CNV or aneuploidy. Accordingly, methods for or determining copy number or ploidy, or detecting CNV or aneuploidy can further include performing a quantitative method to determine copy number or ploidy, or to detect CNV or aneuploidy. The quantitative method can, for example, be the Focal CNV detection using depth of read (FODDOR) classifier method. The method is used for classifying a sample as normal or abnormal. We do this by testing if all the regions of interest, referred to as genes in this discussion of FODDOR, of the sample have the same genetic copy number or different copy numbers. If our test determines that all the genes have the same copy number, we classify the sample as normal. If they have different copy numbers, we classify it as abnormal. Notice that this approach fails to detect abnormal samples that have equal amplifications/deletions in all the regions. The fundamental classifier that we use here is the Generalized likelihood ratio test (GLRT) detector. We frame the problem as follows:

[0248] Let N be the total number of target positions, n_k be the copy number of gene $k \in \{1,...,K\}$, where K is the total number of genes of interest and x_i be the counts at target i, $i \in$

{1,... ,N}, Let g:{1,... ,N} → {1,...,K} be a map from targets to genes. Next, the data is modeled as follows:

$$\log x_i = \log c_s + \log n_{g(i)} + \alpha_s\beta_i + \gamma_s + w_i \qquad (4)$$

where $w_i = K \log \varepsilon_i$ and

$$w_i \sim \mathrm{N}(0; \delta_s^2\, \delta_i^2)$$

. Let $y_i = \log x_i$. Let v_k = log $c_s + \gamma_s + \log(n_k)$ for k E' {1,... ,K}. So, for a healthy gene we have $v_k = \log c_s + \gamma_s + \log(2)$, and for an abnormal gene *k* with a tumor copy number $a_k(\neq 2)$ and tumor fraction *f*, we have $v_k = \log c_s + \gamma_s + \log(2^* (1 - f) + a_k {}^*f)$. Notice that here we are assuming that the whole of gene *k* has the same copy number a_k. In reality the gene may have different copy numbers at different subsections of the gene in which case a_k is the weighted average of the copy numbers of all the subsections of that gene, weighted by the sizes of those subsections. Let us define the virtual tumor fraction of a gene *k* as the amount of excess of that gene compared to the normal genes of that sample, assuming an abnormal copy number of 3 for that gene. So, the virtual tumor fraction is given by

$$vtf_k - (a_k - 2) * f \qquad (5)$$

[0249] For two samples, one with an abnormal copy number of 3 and a tumor fraction f0 and the other with an abnormal copy number of 4 and a tumor fraction of $f_0/2$, the virtual tumor fraction is exactly the same. From the algorithm point of view, these two samples are equivalent. This is because it is theoretically not possible to uniquely determine the abnormal copy number and the true tumor fraction. Also, the *vtf* of a normal gene is zero. Now, if we let $T_s = \log c_s + \gamma_s$ be the sample dependent parameter, we can rewrite the parameter v_k as

$$v_k = T_s + \log(1 + vtf_k) \qquad (6)$$

[0250] For a particular gene *k* with N_k loci, if $y_k = [y_1 : : : y_{Nk}]^T$ is an N_k x 1 vector of logspace normalized depth of reads at the N_k loci and $\beta_k = [\beta_1 : : : \beta_{Nk}]^T$, $\sigma_k = [\sigma_1 : : : \sigma_{Nk}]^T$, $w_k = [w_1 : : : w_{Nk}]^T$, $v_k = [v_1 : : : v_{Nk}]^T$, and define an *N* x *K* matrix U as

$$[\mathbf{U}]_{ik} = \begin{cases} 1 & \text{if } g(i) = k \\ 0 & \text{otherwise} \end{cases}$$

and H = [U β] and $\theta = [v^T\ \alpha_s]^T$. So, in vector form we can rewrite (4) as

$$y = \mathbf{H}\theta + w \qquad (7)$$

where

$$w \sim N(\mathbf{0}, \delta_S^2\, \mathbf{C}(\rho))$$

, $\rho = [\rho_{i} ... \rho_{Nk}]^T$, are the correlation coefficients of each of the genes and $C(\rho) = \mathrm{diag}(C(\rho_1) ... C(\rho_K))$ is a block diagonal matrix where each of its submatrices are as defined in below

$$\mathbf{C}(\rho_k) = (1 - \rho_k) \times \mathbf{diag}(\delta_k^2) + \rho_k \times \sigma_k \sigma_k^T \qquad (8)$$

[0251] Here (β, σ, ρ) are the model parameters which are estimated using known diploid samples as explained in [1]. We can prewhiten the data vector y by multiplying both sides of (7) with S(ρ) where S(ρ) is and N x N matrix such that $S(\rho)^T S(\rho) = C(\rho)^{-1}$. If we let ỹ = S(ρ)y, H̃ = S(ρ)H, and w̃ = S(ρ)w then we can rewrite (4) as

$$\tilde{y} = \tilde{\mathbf{H}}\theta + \tilde{w} \qquad (9)$$

where

$\tilde{w} \sim N(\mathbf{0}, \delta_s^2\, \mathbf{I})$

. Here the unknown parameters are

$\{\theta^T, \delta_s^2\}$

. Let A be a (K -1) x(K +1) "difference matrix" defined as

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -1 & 0 \end{bmatrix} \qquad (10)$$

Notice here that the last column of A is a zero vector which is used to eliminate the nuisance parameter α_s. The hypothesis test we are interested in is

$\mathcal{H}_{0.}$: $A\theta = 0$

$\mathcal{H}_1$: $A\theta \neq 0$

[0252] From Theorem 9.1, as defined by S. M. Kay (see Kay S.M. "Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory". Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1998) the Generalized Likelihood Ratio Test (GLRT) for this hypothesis test is to decide

$\mathcal{H}_1$

. if

$$T(\mathbf{y}) = \frac{N-(K+1)}{K}\,\frac{(\mathbf{A}\hat{\theta}_1 - 0)^T\left[\mathbf{A}(\tilde{\mathbf{H}}^T\tilde{\mathbf{H}})^{-1}\mathbf{A}^T\right]^{-1}(\mathbf{A}\hat{\theta}_1 - 0)}{\tilde{\mathbf{y}}^T(\mathbf{I} - \tilde{\mathbf{H}}(\tilde{\mathbf{H}}^T\tilde{\mathbf{H}})^{-1}\tilde{\mathbf{H}}^T\tilde{\mathbf{y}}} > \gamma' \qquad (11)$$

where is $\hat{\theta}_1 = (\tilde{H}^T\tilde{H})^{-1}\, \tilde{H}^T\tilde{y}$ is the MLE of θ under

$\mathcal{H}_1$

. Notice that the above likelihood ratio is simply a ratio of the sum of squares of multivariate normals due to our assumed noise model. Here *T(y)* is derived starting from the likelihood ratio, and is a monotonically increasing function of it that we have manipulated to turn it into an F-statistic. We assume we that the sample dependent variance (

δ_s^2

) of the noise is unknown and so an MLE for delta is built in to the likelihood function. The exact detection performance (holds for finite data records) is given by

$$P_{FA} = Q_{F_{K,N-(K+1)}}\gamma'$$

$$P_D = Q_{F'_{K,N-(K+1)}(\lambda)}\gamma' \qquad (12)$$

where P_{FA} is the probability of false alarm (false positives), P_D is the probability of detection (true positives), $F_{K,\ N-(K+1)}$ an *F* distribution with *K* numerator degrees of freedom and N-(K + 1) denominator degrees of freedom, and $F'K,_{\ N-(K+1)}(\lambda)$ denotes a noncentral *F* distribution with *K* numerator degrees of freedom, N - (K + 1) denominator degrees of freedom and noncentrality parameter λ. The noncentrality parameter is given by

$$\lambda = \frac{(\mathbf{A}\boldsymbol{\theta}_1 - 0)^T \left[\mathbf{A}(\tilde{\mathbf{H}}^T\tilde{\mathbf{H}})^{-1}\mathbf{A}^T\right]^{-1}(\mathbf{A}\boldsymbol{\theta}_1 - 0)}{\delta_s^2} \qquad (13)$$

where θ_1 is the true value of θ under

$\mathcal{H}_1$

. The Q function is the complement of the cumulative distribution function i.e, $Q(x) = 1 - F_X(x)$. The parameter y' can be set based on the desired performance metrics. For example, we can set the y' based on the desired P_{FA} and the corresponding P_D follows. Note that we cannot simultaneously increase the P_D and decrease the P_{FA} by simply changing the y'.

[0253] In the previous section discussing this FODDOR method, a classifier was designed that at a sample level can classify a sample as normal or abnormal. But that classifier does not tell us which of the genes of an abnormal sample are in deed abnormal. Here we will design a region level classifier which can also determine the individual abnormal genes of an abnormal sample. We do this by iteratively identifying and removing abnormal genes, one per iteration from an abnormal sample, until we find a subset of genes that are normal. Notice that in the previous section, while computing the test statistic, we also estimate the parameter $\theta = [v^T\ \alpha_s]^T$ and so we have an estimate of v. So, for an abnormal sample arg max v should give us the gene with the highest vtf. So, the steps for the iterative region level classifier are as follows:

- Classify a sample as normal or abnormal using the FODDOR classifier. If a sample is normal, then all the regions are normal. If the sample is abnormal go to next step.

- Identify the gene with the highest *vtf* as explained above. Remove this gene from the analysis and go to the previous step.

Notice that this approach has some drawbacks. This approach is less effective when a sample has deletions. When a sample has deletions the algorithm identifies all other regions including normal regions as abnormal and converges to the subset of genes that have deletions and classifies this subset as normal. Accordingly, a method herein can include a quantitative, non-allelic method and a quantitative, allelic method, as provided herein.

Amplification (e.g. PCR) Reaction Mixtures

[0254] Methods of the present invention, in certain embodiments, include forming an amplification reaction mixture. A reaction mixture typically is formed by combining a polymerase, nucleotide triphosphates, nucleic acid fragments from a nucleic acid library generated from the sample, and a set of primer pairs that amplify a set of amplicon that each include a polymorphic loci.. In illustrative embodiments, the reaction mixtures are PCR reaction mixtures. PCR reaction mixtures typically include magnesium.

[0255] In some embodiments, the reaction mixture includes ethylenediaminetetraacetic acid (EDTA), magnesium, tetramethyl ammonium chloride (TMAC), or any combination thereof. In some embodiments, the concentration of TMAC is between 20 and 70 mM, inclusive. While not meant to be bound to any particular theory, it is believed that TMAC binds to DNA, stabilizes duplexes, increases primer specificity, and/or equalizes the melting temperatures of different primers. In some embodiments, TMAC increases the uniformity in the amount of amplified products for the different targets. In some embodiments, the concentration of magnesium (such as magnesium from magnesium chloride) is between 1 and 8 mM.

[0256] The large number of primers used for multiplex PCR of a large number of targets may chelate a lot of the magnesium (2 phosphates in the primers chelate 1 magnesium). For example, if enough primers are used such that the concentration of phosphate from the primers is ~9 mM, then the primers may reduce the effective magnesium concentration by ~4.5 mM. In some embodiments, EDTA is used to decrease the amount of magnesium available as a cofactor for the polymerase since high concentrations of magnesium can result in PCR errors, such as amplification of non-target loci. In some embodiments, the concentration of EDTA reduces the amount of available magnesium to between 1 and 5 mM (such as between 3 and 5 mM).

[0257] In some embodiments, the pH is between 7.5 and 8.5, such as between 7.5 and 8, 8 and 8.3, or 8.3 and 8.5, inclusive. In some embodiments, Tris is used at, for example, a concentration of between 10 and 100 mM, such as between 10 and 25 mM, 25 and 50 mM, 50 and 75 mM, or 25 and 75 mM, inclusive. In some embodiments, any of these concentrations of Tris are used at a pH between 7.5 and 8.5. In some embodiments, a combination of KCl and $(NH_4)_2SO_4$ is used, such as between 50 and 150 mM KCl and between 10 and 90 mM $(NH_4)_2SO_4$, inclusive. In some embodiments, the concentration of KCl is between 0 and 30 mM, between 50 and 100 mM, or between 100 and 150 mM, inclusive. In some embodiments, the concentration of $(NH_4)_2SO_4$ is between 10 and 50 mM, 50 and 90 mM, 10 and 20 mM, 20 and 40 mM, 40 and 60 mM, or 60 and 80 mM $(NH_4)_2SO_4$, inclusive. In some embodiments, the ammonium $[NH_4^+]$ concentration is between 0 and 160 mM, such as between 0 to 50, 50 to 100, or 100 to 160 mM, inclusive. In some embodiments, the sum of the potassium and ammonium concentration ($[K^+] + [NH_4^+]$) is between 0 and 160 mM, such as between 0 to 25, 25 to 50, 50 to 150, 50 to 75, 75 to 100, 100 to 125, or 125 to 160 mM, inclusive. An exemplary buffer with $[K^+] + [NH_4^+] = 120$ mM is 20 mM KCl and 50 mM $(NH_4)_2SO_4$. In some embodiments, the buffer includes 25 to 75 mM Tris, pH 7.2 to 8, 0 to 50 mM KCl, 10 to 80 mM ammonium sulfate, and 3 to 6 mM magnesium, inclusive. In some embodiments, the buffer includes 25 to 75 mM Tris pH 7 to 8.5, 3 to 6 mM $MgCl_2$, 10 to 50 mM KCl, and 20 to 80 mM $(NH_4)_2SO_4$, inclusive. In some embodiments, 100 to 200 Units/mL of polymerase are used. In some embodiments, 100 mM KCl, 50 mM $(NH_4)_2SO_4$, 3 mM $MgCl_2$, 7.5 nM of each primer in the library, 50 mM TMAC, and 7 ul DNA template in a 20 ul final volume at pH 8.1 is used.

[0258] In some embodiments, a crowding agent is used, such as polyethylene glycol (PEG, such as PEG 8,000) or glycerol. In some embodiments, the amount of PEG (such as PEG 8,000) is between 0.1 to 20%, such as between 0.5 to 15%, 1 to 10%, 2 to 8%, or 4 to 8%, inclusive. In some embodiments, the amount of glycerol is between 0.1 to 20%, such as between 0.5 to 15%, 1 to 10%, 2 to 8%, or 4 to 8%, inclusive. In some embodiments, a crowding agent allows either a low polymerase concentration and/or a shorter annealing time to be used. In some embodiments, a crowding agent improves the uniformity of the DOR and/or reduces dropouts (undetected alleles). *Polymerases* In some embodiments, a polymerase with proof-reading activity, a polymerase without (or with negligible) proof-reading activity, or a mixture of a polymerase with proof-reading activity and a polymerase without (or with negligible) proof-reading activity is used. In some embodiments, a hot start polymerase, a non-hot start polymerase, or a mixture of a hot start polymerase and a non-hot start polymerase is used. In some embodiments, a HotStarTaq DNA polymerase is used (see, for example, QIAGEN catalog No. 203203). In some embodiments, AmpliTaq Gold® DNA Polymerase is used. In some embodiments a PrimeSTAR GXL DNA polymerase, a high fidelity polymerase that provides efficient PCR amplification when there is excess template in the reaction mixture, and when amplifying long products, is used (Takara Clontech, Mountain View, CA). In some embodiments, KAPA Taq DNA Polymerase or KAPA Taq HotStart DNA Polymerase is used; they are based on the single-subunit, wild-type *Taq* DNA polymerase of the thermophilic bacterium *Thermus aquaticus.* KAPA Taq and KAPA Taq HotStart DNA Polymerase have 5'-3' polymerase and 5'-3' exonuclease activities, but no 3' to 5' exonuclease (proofreading) activity (see, for example, KAPA BIOSYSTEMS catalog No. BK1000). In some embodiments, *Pfu* DNA polymerase is used; it is a highly thermostable DNA polymerase from the hyperthermophilic archaeum *Pyrococcus furiosus.* The enzyme catalyzes the template-dependent polymerization of nucleotides into duplex DNA in the 5'→3' direction. *Pfu* DNA Polymerase also exhibits 3'→5' exonuclease (proofreading) activity that enables the polymerase to correct nucleotide incorporation errors. It has no 5'→3' exonuclease activity (see, for example, Thermo Scientific catalog No. EP0501). In some embodiments Klentaq1 is used; it is a Klenow-fragment analog of Taq DNA polymerase, it has no exonuclease or endonuclease activity (see, for example, DNA POLYMERASE TECHNOLOGY, Inc, St. Louis, Missouri, catalog No. 100). In some embodiments, the polymerase is a PHUSION DNA polymerase, such as PHUSION High Fidelity DNA polymerase (M0530S, New England BioLabs, Inc.) or PHUSION Hot Start Flex DNA polymerase (M0535S, New England BioLabs, Inc.). In some embodiments, the polymerase is a Q5® DNA Polymerase, such as Q5® High-Fidelity DNA Polymerase (M0491S, New England BioLabs, Inc.) or Q5® Hot Start High-Fidelity DNA Polymerase (M0493S, New England BioLabs, Inc.). In some embodiments, the polymerase is a T4 DNA polymerase (M0203S, New England BioLabs, Inc.).

[0259] In some embodiment, between 5 and 600 Units/mL (Units per 1 mL of reaction volume) of polymerase is used, such as between 5 to 100, 100 to 200, 200 to 300, 300 to 400, 400 to 500, or 500 to 600 Units/mL, inclusive.

PCR Methods

[0260] In some embodiments, hot-start PCR is used to reduce or prevent polymerization prior to PCR thermocycling. Exemplary hot-start PCR methods include initial inhibition of the DNA polymerase, or physical separation of reaction components reaction until the reaction mixture reaches the higher temperatures. In some embodiments, slow release of magnesium is used. DNA polymerase requires magnesium ions for activity, so the magnesium is chemically separated from the reaction by binding to a chemical compound, and is released into the solution only at high temperature. In some embodiments, non-covalent binding of an inhibitor is used. In this method a peptide, antibody, or aptamer are non-covalently bound to the enzyme at low temperature and inhibit its activity. After incubation at elevated temperature, the inhibitor is released and the reaction starts. In some embodiments, a cold-sensitive Taq polymerase is used, such as a modified DNA polymerase with almost no activity at low temperature. In some embodiments, chemical modification is used. In this method, a molecule is covalently bound to the side chain of an amino acid in the active site of the DNA polymerase. The molecule is released from the enzyme by incubation of the reaction mixture at elevated temperature. Once the molecule is released, the enzyme is activated.

[0261] In some embodiments, the amount to template nucleic acids (such as an RNA or DNA sample) is between 20 and 5,000 ng, such as between 20 to 200, 200 to 400, 400 to 600, 600 to 1,000; 1,000 to 1,500; or 2,000 to 3,000 ng, inclusive.

[0262] In some embodiments a QIAGEN Multiplex PCR Kit is used (QIAGEN catalog No. 206143). For 100 x 50 µl multiplex PCR reactions, the kit includes 2x QIAGEN Multiplex PCR Master Mix (providing a final concentration of 3 mM $MgCl_2$, 3 x 0.85 ml), 5x Q-Solution (1 x 2.0 ml), and RNase-Free Water (2 x 1.7 ml). The QIAGEN Multiplex PCR Master Mix (MM) contains a combination of KCl and $(NH_4)_2SO_4$ as well as the PCR additive, Factor MP, which increases the local concentration of primers at the template. Factor MP stabilizes specifically bound primers, allowing efficient primer extension by HotStarTaq DNA Polymerase. HotStarTaq DNA Polymerase is a modified form of *Taq* DNA polymerase and has no polymerase activity at ambient temperatures. In some embodiments, HotStarTaq DNA Polymerase is activated by a 15-minute incubation at 95°C which can be incorporated into any existing thermal-cycler program.

[0263] In some embodiments, 1x QIAGEN MM final concentration (the recommended concentration), 7.5 nM of each primer in the library, 50 mM TMAC, and 7 ul DNA template in a 20 ul final volume is used. In some embodiments, the PCR thermocycling conditions include 95°C for 10 minutes (hot start); 20 cycles of 96°C for 30 seconds; 65°C for 15 minutes; and 72°C for 30 seconds; followed by 72°C for 2 minutes (final extension); and then a 4°C hold.

[0264] In some embodiments, 2x QIAGEN MM final concentration (twice the recommended concentration), 2 nM of each primer in the library, 70 mM TMAC, and 7 ul DNA template in a 20 ul total volume is used. In some embodiments, up to 4 mM EDTA is also included. In some embodiments, the PCR thermocycling conditions include 95°C for 10 minutes (hot start); 25

cycles of 96°C for 30 seconds; 65°C for 20, 25, 30, 45, 60, 120, or 180 minutes; and optionally 72°C for 30 seconds); followed by 72°C for 2 minutes (final extension); and then a 4°C hold.

[0265] Another exemplary set of conditions includes a semi-nested PCR approach. The first PCR reaction uses 20 ul a reaction volume with 2x QIAGEN MM final concentration, 1-5 nM of each primer in the library, and DNA template.

[0266] Thermocycling parameters include 95°C for 10 minutes; 25 cycles of 96°C for 30 seconds, 65°C for 1 minute, 58°C for 6 minutes, 60°C for 8 minutes, 65°C for 4 minutes, and 72°C for 30 seconds; and then 72°C for 2 minutes, and then a 4°C hold. Next, 2 ul of the resulting product, diluted 1:200, is used as input in a second PCR reaction. This reaction can include, for example, a 10 ul reaction volume with 1x QIAGEN MM final concentration, 20 nM of each primer of a set of primer pairs. Thermocycling parameters can include, for example, 95°C for 10 minutes; 15 cycles of 95°C for 30 seconds, 65°C for 1 minute, 60°C for 5 minutes, 65°C for 5 minutes, and 72°C for 30 seconds; and then 72°C for 2 minutes, and then a 4°C hold. The annealing temperature can optionally be higher than the melting temperatures of some or all of the primers, as discussed herein.

[0267] The melting temperature (T_m) is the temperature at which one-half (50%) of a DNA duplex of an oligonucleotide (such as a primer) and its perfect complement dissociates and becomes single strand DNA. The annealing temperature (T_A) is the temperature one runs the PCR protocol at. For prior methods, it is usually 5°C below the lowest T_m of the primers used, thus close to all possible duplexes are formed (such that essentially all the primer molecules bind the template nucleic acid). While this is highly efficient, at lower temperatures there are more unspecific reactions bound to occur. One consequence of having too low a T_A is that primers may anneal to sequences other than the true target, as internal single-base mismatches or partial annealing can be tolerated. In some embodiments of the present inventions, the T_A is higher than T_m, where at a given moment only a small fraction of the targets have a primer annealed (such as only -1-5%). If these get extended, they are removed from the equilibrium of annealing and dissociating primers and target (as extension increases T_m quickly to above 70°C), and a new -1-5% of targets has primers. Thus, by giving the reaction a long time for annealing, one can get -100% of the targets copied per cycle.

[0268] In various embodiments, the annealing temperature is between 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 °C and 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, or 15 °C on the high end of the range, greater than the melting temperature (such as the empirically measured or calculated T_m) of at least 25, 50, 60, 70, 75, 80, 90, 95, or 100% of the non-identical primers. In various embodiments, the annealing temperature is between 1 and 15 °C (such as between 1 to 10, 1 to 5, 1 to 3, 3 to 5, 5 to 10, 5 to 8, 8 to 10, 10 to 12, or 12 to 15 °C, inclusive) greater than the melting temperature (such as the empirically measured or calculated T_m) of at least 25; 50; 75; 100; 300; 500; 750; 1,000; 2,000; 5,000; 7,500; 10,000; 15,000; 19,000; 20,000; 25,000; 27,000; 28,000; 30,000; 40,000; 50,000; 75,000; 100,000; or all of the non-identical primers. In various embodiments, the annealing temperature is between 1 and 15 °C (such as between 1 to 10, 1 to 5, 1 to 3, 3 to 5, 3 to 8, 5 to 10, 5 to 8, 8 to 10, 10 to 12, or 12 to 15 °C, inclusive) greater than the melting temperature (such as the empirically measured or calculated T_m) of at least 25%, 50%, 60%, 70%, 75%, 80%, 90%, 95%, or all of the non-identical primers, and the length of the annealing step (per PCR cycle) is between 5 and 180 minutes, such as 15 and 120 minutes, 15 and 60 minutes, 15 and 45 minutes, or 20 and 60 minutes, inclusive.

Exemplary Multiplex PCR Methods

[0269] In various embodiments, limiting primer concentrations and/or conditions are used. In various embodiments, the length of the annealing step is between 15, 20, 25, 30, 35, 40, 45, or 60 minutes on the low end of the range and 20, 25, 30, 35, 40, 45, 60, 120, or 180 minutes on the high end of the range. In various embodiments, the length of the annealing step (per PCR cycle) is between 30 and 180 minutes. For example, the annealing step can be between 30 and 60 minutes and the concentration of each primer can be less than 20, 15, 10, or 5 nM. In other embodiments the primer concentration is 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, or 25 nM on the low end of the range, and 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, and 50 on the high end of the range.

[0270] At high level of multiplexing, the solution can become viscous due to the large amount of primers in solution. If the solution is too viscous, one can reduce the primer concentration to an amount that is still sufficient for the primers to bind the template DNA. In various embodiments, between 500 and 100,000 different primers are used and the concentration of each primer is less than 20 nM, such as less than 10 nM or between 1 and 10 nM, inclusive.

Example Computer Architecture

[0271] In certain embodiments, which are not claimed, provided herein are computer programs and computer systems for performing the analytical steps of the methods provided herein, such as estimating the phase, generating individual probabilities, generating joint probabilities, generating a set of hypothesis or models, and/or selecting a best fit model, using genetic data generated using the kit. The computer programs in certain embodiments, are associated with pools, sets, pluralities, or libraries of primers as provided herein, for carrying out methods provided herein.

[0272] In some embodiments, provided herein is a system for detecting chromosomal ploidy in a sample of an individual. The system can include the following:

1. a. an input processor configured to receive allelic frequency data comprising the amount of each allele present in the sample at each loci of a plurality of polymorphic loci, for example a set of SNP loci, on a plurality of segments within the chromosomal region, wherein each segment comprises loci with strong linkage disequilibrium, or each segment is a haploblock;

2. b. a modeler configured to:
    1. i. generate phased allelic information for the set of polymorphic loci by estimating the phase of the allele frequency data taking into account an increased statistical correlation of polymorphic loci within the same segment;
    2. ii. generate individual probabilities of allele frequencies for the polymorphic loci for different ploidy states using the allele frequency data; and
    3. iii. generate joint probabilities for the set of polymorphic loci using the individual probabilities and the phased allelic information; and
3. c. a hypothesis manager configured to select, based on the joint probabilities, a best fit model indicative of chromosomal ploidy, thereby determining ploidy of the chromosomal region.

[0273] In certain system embodiments, the allele frequency data is generated by a nucleic acid sequencing system.

[0274] A nontransitory computer readable medium for detecting chromosomal ploidy in a sample of an individual, when executed by a processing device, may cause the processing device to perform the following:

1. a. receive allele frequency data comprising the amount of each allele present in the sample at each loci of a plurality of polymorphic loci on a plurality of segments within the chromosomal region, wherein each segment comprises loci with strong linkage disequilibrium;
2. b. generate phased allelic information for the set of polymorphic loci by estimating the phase of the allele frequency data taking into account an increased statistical correlation of polymorphic loci within the same segment;
3. c. generate individual probabilities of allele frequencies for the polymorphic loci for different ploidy states using the allele frequency data;
4. d. generate joint probabilities for the set of polymorphic loci using the individual probabilities and the phased allelic information; and
5. e. select, based on the joint probabilities, a best fit model indicative of chromosomal ploidy, thereby determining ploidy of the chromosomal region.

[0275] The allele frequency data may be generated from nucleic acid sequence data.

[0276] FIG. 5 shows an example system architecture X00 useful for performing embodiments of the present invention. System architecture X00 includes an analysis platform X08 connected to one or more laboratory information systems ("LISs") X04. As shown in FIG. 5, analysis platform X08 can be connected to LIS X04 over a network X02. Network X02 may include one or more networks of one or more network types, including any combination of LAN, WAN, the Internet, etc. Network X02 may encompass connections between any or all components in system architecture X00. Analysis platform X08 may alternatively or additionally be connected directly to LIS X06. In an embodiment, analysis platform X08 analyzes genetic data provided by LIS X04 in a software-as-a-service model, where LIS X04 is a third-party LIS, while analysis platform X08 analyzes genetic data provided by LIS X06 in a full-service or in-house model, where LIS X06 and analysis platform X08 are controlled by the same party. In an embodiment where analysis platform X08 is providing information over network X02, analysis platform X08 can be a server.

[0277] In an example embodiment, laboratory information system X04 includes one or more public or private institutions that collect, manage, and/or store genetic data. A person having skill in the relevant art(s) would understand that methods and standards for securing genetic data are known and can be implemented using various information security techniques and policies, e.g., username/password, Transport Layer Security (TLS), Secure Sockets Layer (SSL), and/or other cryptographic protocols providing communication security.

[0278] In an example embodiment, system architecture X00 operates as a service-oriented architecture and uses a client-server model that would be understood by one of skill in the relevant art(s) to enable various forms of interaction and communication between LIS X04 and analysis platform X08. System architecture X00 can be distributed over various types of networks X02 and/or may operate as cloud computing architecture. Cloud computing architecture may include any type of distributed network architecture. By way of example and not of limitation, cloud computing architecture is useful for providing software as a service (SaaS), infrastructure as a service (IaaS), platform as a service (PaaS), network as a service (NaaS), data as a service (DaaS), database as a service (DBaaS), backend as a service (BaaS), test environment as a service (TEaaS), API as a service (APIaaS), integration platform as a service (IPaaS) etc.

[0279] In an example, LISs X04 and X06 each include a computer, device, interface, etc. or any sub-system thereof. LISs X04 and X06 may include an operating system (OS), applications installed to perform various functions such as, for example, access to and/or navigation of data made accessible locally, in memory, and/or over network X02. In an embodiment, LIS X04 accesses analysis platform X08 through an application programming interface ("API"). LIS X04 may also include one or more native applications that may operate independently of an API.

[0280] In an example, analysis platform X08 includes one or more of an input processor X12, a hypothesis manager X14, a modeler X16, an error correction unit X18, a machine learning unit X20, and an output processor X18. Input processor X12 receives and processes inputs from LISs X04 and/or X06. Processing may include but is not limited to operations such as parsing, transcoding, translating, adapting, or otherwise handling any input received from LISs X04 and/or X06. Inputs can be received via one or more streams, feeds, databases, or other sources of data, such as can be made accessible by LISs X04 and X06. Data errors can be corrected by error correction unit X18 through performance of the error correction mechanisms described above.

[0281] In an example, hypothesis manager X14 is configured to receive the inputs passed from input processor X12 in a form ready to be processed in accordance with hypotheses for genetic analysis that are represented as models and/or algorithms. Such models and/or algorithms can be used by modeler X16 to generate probabilities, for example, based on dynamic, real-time, and/or historical statistics or other indicators. Data used to derive and populate such strategy models and/or algorithms are available to hypothesis manager X14 via, for example, genetic data source X10. Genetic data source X10 may include, for example, a nucleic acid sequencer. Hypothesis manager X14 can be configured to formulate hypotheses based on, for example, the variables required to populate its models and/or algorithms. Models and/or algorithms, once populated, can be used by modeler X16 to generate one or more hypotheses as described above. Hypothesis manager X14 may select a particular value, range of values, or estimate based on a most-likely hypothesis as an output as described above. Modeler X16 may operate in accordance with models and/or algorithms trained by machine learning unit X20. For example, machine learning unit X20 may develop such models and/or algorithms by applying a classification algorithm as described above to a training set database (not shown).

[0282] Once hypothesis manager X14 has identified a particular output, such output can be returned to the particular LIS 104 or 106 requesting the information by output processor X22.

[0283] Various aspects of the disclosure can be implemented on a computing device by software, firmware, hardware, or a combination thereof. FIG. 6 illustrates an example computer system Y00 in which the contemplated embodiments, or portions thereof, can be implemented as computer-readable code. Various embodiments are described in terms of this example computer system Y00.

[0284] Processing tasks in the embodiment of FIG. 6 are carried out by one or more processors Y02. However, it should be noted that various types of processing technology can be used here, including programmable logic arrays (PLAs), application-specific integrated circuits (ASICs), multicore processors, multiple processors, or distributed processors. Additional specialized processing resources such as graphics, multimedia, or mathematical processing capabilities may also be used to aid in certain processing tasks. These processing resources can be hardware, software, or an appropriate combination thereof. For example, one or more of processors Y02 can be a graphics-processing unit (GPU). In an embodiment, a GPU is a processor that is a specialized electronic circuit designed to rapidly process mathematically intensive applications on electronic devices. The GPU may have a highly parallel structure that is efficient for parallel processing of large blocks of data, such as mathematically intensive data. Alternatively or in addition, one or more of processors Y02 can be a special parallel processing without the graphics optimization, such parallel processors performing the mathematically intensive functions described herein. One or more of processors Y02 may include a processing accelerator (e.g., DSP or other special-purpose processor).

[0285] Computer system Y00 also includes a main memory Y30, and may also include a secondary memory Y40. Main memory Y30 can be a volatile memory or non-volatile memory, and divided into channels. Secondary memory Y40 may include, for example, non-volatile memory such as a hard disk drive Y50, a removable storage drive Y60, and/or a memory stick. Removable storage drive Y60 may comprise a floppy disk drive, a magnetic tape drive, an optical disk drive, a flash memory, or the like. The removable storage drive Y60 reads from and/or writes to a removable storage unit 470 in a well-known manner. Removable storage unit Y70 may comprise a floppy disk, magnetic tape, optical disk, etc. which is read by and written to by removable storage drive Y60. As will be appreciated by persons skilled in the relevant art(s), removable storage unit Y70 includes a computer usable storage medium having stored therein computer software and/or data.

[0286] In alternative implementations, secondary memory Y40 may include other similar means for allowing computer programs or other instructions to be loaded into computer system Y00. Such means may include, for example, a removable storage unit Y70 and an interface (not shown). Examples of such means may include a program cartridge and cartridge interface (such as that found in video game devices), a removable memory chip (such as an EPROM, or PROM) and associated socket, and other removable storage units Y70 and interfaces which allow software and data to be transferred from the removable storage unit Y70 to computer system Y00.

[0287] Computer system Y00 may also include a memory controller Y75. Memory controller Y75 controls data access to main memory Y30 and secondary memory Y40. In some embodiments, memory controller Y75 can be external to processor Y10, as shown in FIG. 6. In other embodiments, memory controller Y75 may also be directly part of processor Y10. For example, many AMDTM and IntelTM processors use integrated memory controllers that are part of the same chip as processor Y10 (not shown in FIG. 6).

[0288] Computer system Y00 may also include a communications and network interface Y80. Communication and network interface Y80 allows software and data to be transferred between computer system Y00 and external devices. Communications and network interface Y80 may include a modem, a communications port, a PCMCIA slot and card, or the like. Software and data transferred via communications and network interface Y80 are in the form of signals which can be electronic, electromagnetic, optical, or other signals capable of being received by communication and network interface Y80. These signals are provided to communication and network interface Y80 via a communication path Y85. Communication path Y85 carries signals and can be implemented using wire or cable, fiber optics, a phone line, a cellular phone link, an RF link or other communications channels.

[0289] The communication and network interface Y80 allows the computer system Y00 to communicate over communication networks or mediums such as LANs, WANs the Internet, etc. The communication and network interface Y80 may interface with remote sites or networks via wired or wireless connections.

[0290] In this document, the terms "computer program medium," "computer-usable medium" and "non-transitory medium" are used to generally refer to tangible media such as removable storage unit Y70, removable storage drive Y60, and a hard disk installed in hard disk drive Y50. Signals carried over communication path Y85 can also embody the logic described herein. Computer program medium and computer usable medium can also refer to memories, such as main memory Y30 and secondary memory Y40, which can be memory semiconductors (e.g. DRAMs, etc.). These computer program products are means for providing software to computer system Y00.

[0291] Computer programs (also called computer control logic) are stored in main memory Y30 and/or secondary memory Y40. Computer programs may also be received via communication and network interface Y80. Such computer programs, when executed, enable computer system Y00 to implement embodiments as discussed herein. In particular, the computer programs, when executed, enable processor Y10 to implement the disclosed processes. Accordingly, such computer programs represent controllers of the computer system Y00. Where the embodiments are implemented using software, the software can be stored in a computer program product and loaded into computer system Y00 using removable storage drive Y60, interfaces, hard drive Y50 or communication and network interface Y80, for example.

[0292] The computer system Y00 may also include input/output/display devices Y90, such as keyboards, monitors, pointing devices, touchscreens, etc.

[0293] It should be noted that the simulation, synthesis and/or manufacture of various embodiments can be accomplished, in part, through the use of computer readable code, including general programming languages (such as C or C++), hardware description languages (HDL) such as, for example, Verilog HDL, VHDL, Altera HDL (AHDL), or other available programming tools. This computer readable code can be disposed in any known computer-usable medium including a semiconductor, magnetic disk, optical disk (such as CD-ROM, DVD-ROM). As such, the code can be transmitted over communication networks including the Internet.

[0294] The embodiments are also directed to computer program products comprising software stored on any computer-usable medium. Such software, when executed in one or more data processing devices, causes a data processing device(s) to operate as described herein. Embodiments employ any computer-usable or -readable medium, and any computer-usable or -readable storage medium known now or in the future. Examples of computer-usable or computer-readable mediums include, but are not limited to, primary storage devices (e.g., any type of random access memory), secondary storage devices (e.g., hard drives, floppy disks, CD ROMS, ZIP disks, tapes, magnetic storage devices, optical storage devices, MEMS, nano-technological storage devices, etc.), and communication mediums (e.g., wired and wireless communications networks, local area networks, wide area networks, intranets, etc.). Computer-usable or computer-readable mediums can include any form of transitory (which include signals) or non-transitory media (which exclude signals). Non-transitory media comprise, by way of non-limiting example, the aforementioned physical storage devices (e.g., primary and secondary storage devices).

[0295] The features disclosed in the foregoing description, or the following claims, or the accompanying drawings, expressed in their specific forms or in terms of a means for performing the disclosed function, or a method or process for attaining the disclosed result, as appropriate, may, separately, or in any combination of such features, be utilised for realising the invention as defined in the claims.

[0296] The following examples are put forth so as to provide those of ordinary skill in the art with a complete disclosure and description of how to use the embodiments provided herein, and are not intended to limit the scope of the disclosure nor are they intended to represent that the Examples below are all or the only experiments performed. Efforts have been made to ensure accuracy with respect to numbers used (e.g. amounts, temperature, etc.) but some experimental errors and deviations should be accounted for. Unless indicated otherwise, parts are parts by volume, and temperature is in degrees Centigrade. It should be understood that variations in the methods as described can be made without changing the fundamental aspects that the Examples are meant to illustrate.

EXAMPLES

Example 1. Creation of primer pool for ovarian cancer polymorphic loci within haploblocks

[0297] This example illustrates a method for identifying haploblocks within target chromosomal regions for detecting CNV in ovarian cancer, identifying target polymorphic loci within those segments, and selecting a pool of primers for amplifying nucleic acids including those target polymorphic loci. to the pool of primers allow the determination of allele frequencies at those polymorphic loci in experiments provided in other Examples herein. Accordingly, in this example, Ovarian cancer chromosome regions of interest were identified, haploblocks were identified, candidate SNPs were selected, and pools of primers were designed for amplifying the candidate SNPs.

Primer Pool Design.

[0298] The design process consists of these main steps:

1. a. Select candidate target SNPs for each region of interest.
2. b. Attempt to design up to five sets of right and left specific primers for each candidate target SNP.
3. c. Filter designs into haploblocks with at least 10 SNPs with designs.

4. d. Select compatible designs to form the primer pools.

Candidate SNPs Selection:

[0299] For each region of interest, we chose candidate SNPs satisfying the following criteria:
e. The SNP must be present in both dbSNP Common 138 and the 1000 Genomes project (the phase 1 version 3 variant calls released April 30, 2012, "An integrated map of genetic variation from 1,092 human genomes," McVean et al, Nature 491: 56-65 (01 November 2012) doi:10.1038/nature11632) variant call data set.
f. The SNP minor allele frequency from the 1000 Genomes project must be at least 10%.
g. The SNP location must be within one of the corresponding breakpoints in Table 4.

Table 4:

Chrom	Start	End	Event Type	No.Patients (COSMIC)	Cancer Census Gene
8	115298000	145233000	GAIN	173	MYC, MTSS1, NDRG1
3	166356000	180256000	GAIN	108	PIK3CA, MECOM
8	100758000	115298000	GAIN	101	
8	617000	37343000	LOSS	99	
19	28240000	33433000	GAIN	82	CCNE1
*20	29369569	63025520	GAIN	82	
*20	1	26369569	GAIN	67	
12	18959000	29050000	GAIN	65	KRAS
19	34341000	40857000	GAIN	55	AKT2
19	12042000	17796000	GAIN	54	
16	60437000	89380000	LOSS	50	CDH1
*17	25800001	31800000	LOSS	30	NF1
22	42378000	49332000	LOSS	21	
*17	10700001	16000000	LOSS	16	MAP2K4

Table 5: Number of candidate SNPs selected for each region of interest.

Chrom	Start	End	Candidate SNPs
8	115298000	145233000	61,362
3	166356000	180256000	24,023
8	100758000	115298000	25,035
8	617000	37343000	96,572
19	28240000	33433000	10,294

Chrom	Start	End	Candidate SNPs
*20	29369569	63025520	60,135
*20	1	26369569	54,321
12	18959000	29050000	19,888
19	34341000	40857000	11,607
19	12042000	17796000	12,303
16	60437000	89380000	66,790
*17	25800001	31800000	7,699
22	42378000	49332000	17,705
*17	10700001	16000000	12,111

Primer design:

[0300] Primers were designed using primer3 release 2.3.6 (Whitehead Institute for Biomedical Research, Steve Rozen (Available on the Internet at //primer3.sourceforge.net/releases.php)) and then filtered in a reiterative fashion to check primer specificity. For each candidate SNP primer3 was used to design left and right primers (two-sided) with an amplicon length within a range of 50 to 75 bp and a melting temperature between 53-60°C. Primer3 was configured to use the SantaLucia salt correction and melting temperature formulae (SantaLucia JR (1998) "A unified view of polymer, dumbbell and oligonucleotide DNA nearest-neighbor thermodynamics", Proc Natl Acad Sci 95:1460-65).

[0301] Primer locations are restricted to be at least 2bp away from any SNP which is present either in dbSNP Common 138, or in the 1000 Genomes project with minor allele frequency larger than 1%. Up to five designs can be generated per target. The parameters in Table 6 were used for primer design.

Table 6. Exemplary design parameters:

Name	Original Value	Description
target_padding	2	Primers should end at least 2 bases away from the target loci
min_amplicon_size	50	
max_amplicon_size	75	
PRIMER_MAX_SIZE	30	
PRIMER_OPT_SIZE	24	
PRIMER_MIN_SIZE	18	
PRIMER_WT_SIZE_LT	0	

Name	Original Value	Description
PRIMER_WT_SIZE_GT	1	Penalty for primer longer than optimal
PRIMER_PAIR_WT_PRODUCT_SIZE_LT	0	
PRIMER_PAIR_WT_PRODUCT_SIZE_GT	3	Significant penalty for amplicon longer than optimal
PRIMER_MAX_TM	60	
PRIMER_OPT_TM	56	
PRIMER_MIN_TM	53	
PRIMER_WT_TM_LT	1.5	Penalty for TM lower than optimal
PRIMER_WT_TM_GT	1	Penalty for TM higher than optimal
PRIMER_MAX_GC	70	
PRIMER_OPT_GC_PERCENT	50	
PRIMER_MIN_GC	30	
PRIMER_WT_GC_PERCENT_LT	1	
PRIMER_WT_GC_PERCENT_GT	1	
PRIMER_MAX_END_GC	3	
PRIMER_SALT_CORRECTIONS	1	
PRIMER_MAX_POLY_X	10	
PRIMER_INTERNAL_MAX_POLY_X	10	

[0302] The designs generated by primer3 were then filtered:

h. if the amplicon GCcontent is not in a safe range [30%-70%].

i. if primer pairs are susceptible to mispriming and amplicons that are not sufficiently unique in the genome to map confidently.

[0303] Finally, for SNPs with multiple remaining design pairs we keep the shortest amplicon. The following table shows the number of SNPs with passing designs. It should be noted that many if not most candidate SNPs do not have any feasible design.

Table 7: Number of SNPs with designed assays for each region of interest.

Chrom	**Start**	**End**	**SNPs with design**	**Yield**
8	115298000	145233000	15,993	26.1%

Chrom	Start	End	SNPs with design	Yield
3	166356000	180256000	4,194	17.5%
8	100758000	115298000	4,644	18.6%
8	617000	37343000	16,503	17.1%
19	28240000	33433000	3,041	29.5%
*20	29369569	63025520	17,371	28.9%
*20	1	26369569	12,955	23.8%
12	18959000	29050000	3,289	16.5%
19	34341000	40857000	2,649	22.8%
19	12042000	17796000	2,510	20.4%
16	60437000	89380000	15,082	22.6%
*17	25800001	31800000	1,842	23.9%
13	48765000	49720000	230	19.5%
22	42378000	49332000	5,756	32.5%
*17	10700001	16000000	2,541	21.0%

[0304] Haploblocks were identified by identifying polymorphic loci with strong linkage disequilibrium using a D' > 95% cutoff where 95% of pairwise SNP comparisons showed a strong linkage disequilibrium. SNPs with minor allele frequency of less than 5% were ignored by the method. The program called plink was used to estimate haploblocks (http://pngu.mgh.harvard.edu/~purcell/plink/ld.shtml#blox). The program estimates haploblocks for a given set of SNPs based on a given reference panel.

[0305] For wet lab experiments confirming the in silico results, amplicons can be identified that include the SNPs, with lengths between 50 and 75 bp, with a Tm of between 53-66C and with a GC content of 30-70 and MAF of 10-50%.

[0306] We used the 1000 genomes project haplotypes as the reference panel (1000 genomes project haplotypes release September 2013). The release contains haplotypes on 1092 samples (#haplotypes = 2184) for 36.8 million SNPs. The haploblocks in Table 8 were identified.

Table 8: Identified haploblocks for SNPs with designed assays for each region of interest

Index	**Chrom**	**start_bp**	**end_bp**	designs in haploblocks >-10	Yield	designs in haploblocks > 20	**Longest _Block**
1	12	18,959,000	29,050,000	645	77%	221	42
3	16	60,437,000	89,380,000	1170	95%	670	44
4	19	12,042,000	17,796,000	405	72%	104	29
5	19	28,240,000	33,433,000	836	82%	343	61

Index	**Chrom**	**start_bp**	**end_bp**	designs in haploblocks >-10	Yield	designs in haploblocks > 20	**Longest _Block**
6	19	34,341,000	40,857,000	704	84%	402	42
7	22	42,378,000	49,332,000	547	70%	156	57
8	3	166,356,000	180,256,000	771	78%	266	37
9	8	617,000	37,343,000	1225	92%	624	55
10	8	115,298,000	145,233,000	1173	91%	708	57
11	8	100,758,000	115,298,000	1309	97%	628	64
12	20	1	26,369,569	1016	96%	769	77
13	20	29,369,569	63,025,520	1238	96%	965	50
14	17	25,800,001	31,800,000	457	78%	173	24
15	17	10,700,001	16,000,000	332	84%	126	24

Pooling:

[0307] Candidate PCR assays are ranked and selected on the basis of number of patients having a CNV spanning over the SNP location, the haploblock size in terms of number of SNPs with haploblocks with more SNPs being favored, target SNP minor allele frequency, observed heterozygosity rate (from dbSNP), presence in HapMap, type of mutation (transversions are preferred over transitions), amplicon GC-content and amplicon length.

Results

[0308] Table 9 provides details regarding haploblocks (i.e. target segments) within chromosome regions of interest. As indicated, for each segment there were at least 81 SNPs in haploblocks (i.e. segments) with greater than 20 SNPs. The longest haploblock per chromosome region of interest varied from 24 to 79 SNPs (Table 9).

Table 9: Final pool configuration for each region of interest.

Chrom	**Start**	**End**	Number of assays	expected no .hets	**Number of SNPs in Blocks >20**	**Longest block**
8	115298000	145233000	1296	507	829	56
3	166356000	180256000	992	365	290	40
8	100758000	115298000	1354	512	729	61
8	617000	37343000	1332	512	608	55
19	28240000	33433000	1019	388	336	54

Chrom	**Start**	**End**	Number of assays	expected no .hets	**Number of SNPs in Blocks >20**	**Longest block**
*20	29369569	63025520	1290	508	1041	50
*20	1	26369569	1055	406	790	79
12	18959000	29050000	843	294	221	42
19	34341000	40857000	838	317	393	42
19	12042000	17796000	563	208	81	30
16	60437000	89380000	1235	453	720	43
*17	25800001	31800000	588	230	174	24
22	42378000	49332000	783	307	201	57
*17	10700001	16000000	395	144	125	24

Example 2

[0309] In this in silico experiment the accuracy of the informatics haplotyping was determined. To estimate haplotypes, the tool ShapeIt was used (available at the hypertext transfer protocol secure site at mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html). ShapeIt takes as input a list of genotypes along with haplotyping likelihoods based on SNP loci locations and population cross-over data, and outputs estimated haplotypes for the inputted genotypes. It estimates haplotypes for each chromosome separately.

[0310] The 1000 genomes project has existing high confidence genotype calls for many individuals publicly available. The entirety of this high quality genotype dataset was used as a test dataset for the haplotyping validation. Similarly, the 1000 genomes project has available high quality haplotyping information for each dataset. The 1000 genomes haplotyping data can be used as a best guess truth dataset for comparison.

[0311] Comparing haplotypes estimated by ShapeIt to known, curated haplotypes from 1000 genomes provides us with a measure of the level of haplotyping accuracy and error of the primer pool. When comparing errors in haplotyping, it is important to also consider if the mis-haplotyping is occurring within a known haploblock or outside of a haploblock. SNPs within a haploblock are genetically linked and generally exist together. Thus, one can conclude that the mis-haplotyping switch error will be lesser within haploblocks and greater outside of haploblocks.

[0312] 1092 genotype samples from the 1000 genomes dataset were used for the validation. All samples were run through ShapeIt for haplotype estimation. The resulting haplotypes were compared to existing, curated 1000 genomes haplotypes to determine the level of error in haplotyping for the primer pool set. Each haplotyping event was carried out on each region independently.

[0313] The Haplotyping Error is calculated as:
Haplotyping Error = (Number of switched haplotype calls at SNP X)/(Number of heterozygous genotypes). It was observed that haplotying error rates were decreased within haplblocks in simulations.

Example 3

[0314] In this in silico experiment, it was observed that by analyzing polymorphic data as if the polymorphic loci were within haploblocks, allelic imbalance was detected at similar rates to calculations using perfect haplotype data, in samples down to allelic imbalances of 1%, provided that a sufficient number of polymorphic loci per target chromosome region were analyzed that were within haploblocks having a minimum number of polymorphic loci per haploblock.. Two artificial titration experiments using breast cancer cell lines (HCC1954 and HCC2218) were performed to evaluate the performance of the CNV calling algorithm in plasma samples. More specifically, titrations were prepared from pairs of matched tumor and normal cell line samples and having CNVs on chromosome 1 or chromosome 2.

[0315] Cell line HCC1954 was evaluated for chromosome 1, and cell line HCC2218 was evaluated for chromosome 2. For each chromosome, 1248 SNPs were analyzed.

[0316] We assigned certain numbers of consecutive SNPs to haploblocks to evaluate the theoretical performance of the CNV calling algorithm in the potential product. The allele count data from published titration experiments (Kirkizlar et al. 2015 (Kirkizlar et al., Translational Oncology, 8 407-416)) were used. The probes of Kirkizlar et al. 2015 were used, except that if there were more than 1248 SNPs in a probe design, only the first 1248 SNPs were used.

[0317] We assumed that we had perfect haplotype information within the blocks, and no haplotype information between the blocks. Referring to the formula provided herein in the section on combining likelihoods, (Combined_Likelihoods), in the presence of perfect haplotype information, we have c=0 or c=1. In the present simulation to determine the optimal block size, we assumed perfect haplotypes within the blocks (i.e., c=0 or c=1) and we assumed no haplotype information between the blocks (i.e., c=0.5). Note that as the minimum block size increases, the number of total SNPs decreases. We attempted to determine the optimal minimum block size that also has a sufficient number of SNPs. We ran our algorithm for minimum block sizes of 1, 10, 15 and 20; and compared our results with the system that had perfect haplotype information.

[0318] For minimum block size of 1 (i.e., no haploblock requirement), especially regions with very few good blocks had false positives (with > 1.0% allelic imbalance detected for multiple regions that were negative). The quantification of the allelic imbalance value became more accurate for >2.0% allelic imbalance.

[0319] Performance of the algorithm was similar to the perfect haplotype case for minimum block size of 10 and maximum block size of 100, and sufficient number of SNPs (≥ 1000). More specifically, for such cases, there have been scenarios with false positives (allelic imbalance of >0.50%), but generally the detection of true positives has been successful (for each case with ≥ 1000 SNPs in haploblocks that had allelic imbalance of >1.0% originally, the allelic imbalance was detected to be >1.0% in the imperfect haplotype scenario).

[0320] However, for scenarios with a low number of SNPs (i.e. 125 to 250), the algorithm failed to detect even allelic imbalance of >2.5%. Hence, a minimum block size of 10 and at least 350 SNPs in each region, proved to be especially effective for the Ovarian cancer arm length CNV analysis performed in this simulation. Note that for other cancers and for focal chromosome regions, smaller numbers of SNPs and smaller minimum number of SNPs per haploblock can be successfully employed (See Example 5 - lung cancer focal chromosome region analysis).

Example 4

[0321] This example confirms the effectiveness of the methods provided herein, particularly methods that include the haploblock assay/primer design step of Example 1, in a wet lab environment with patient samples, biochemical methods. Accordingly, for this experiment the primers/assays for ovarian cancer identified in Example 1, were used.

Sample preparation

[0322] DNA extraction and QC. All the plasma aliquots from each patient were pooled prior to cfDNA extraction, and the hemolysis grade of each pooled plasma sample was evaluated visually (no hemolysis, mild hemolysis or severe hemolysis). cfDNA was extracted using the Qiagen NA kit (Valencia, CA) following a protocol optimized for 5 ml of plasma. All cfDNA samples were QCed on Bioanalyzer High Sensitivity chips (Agilent, Santa Clara, CA). The same Bioanalyzer High Sensitivity runs were also used to quantify the cfDNA samples by interpolation of the mononucleosomal peak height on a calibration curve prepared from a pure cfDNA sample that was previously quantified. This was necessary because cfDNA sometimes contains an intact DNA fraction that overlaps with the high size marker on the chip, which makes quantification of the mononucleosomal peak unreliable. A representative subset of the purified genomic DNA samples was quantified using Nanodrops (Wilmington, DE). All of the samples quantified were in the expected range (~10 ng/μl).

[0323] cfDNA library preparation. The entire cfDNA amount from each plasma sample was used as input into Library Prep using the Natera library prep kit and following the kit instructions. Libraries were generated from the samples above. Adapters were ligated to DNA fragments and the fragments were amplified using the following protocol:

[0324] 95°C, 2 min; 15 x [95°C, 20 sec, 55°C, 20 sec, 68°C, 20 sec], 68°C 2 min, 4°C hold. The libraries were amplified to plateau and then purified using Ampure beads (Beckman Coulter, Brea, CA) following the manufacturer's protocol. The purified libraries were QCed on the LabChip.

[0325] cfDNA multiplex PCR and Sequencing. The library material from each plasma sample was used as input DNA into multiplex PCR (mPCR) reactions in the relevant assay pool and an optimized plasma mPCR protocol. The primers of Table 9 of Example 1 were obtained (IDT, Coralville, Iowa) as a pool. A 10nM primer concentration was used for each primer. The reactions were performed using the following protocol: PCR amplified: 95C 10min, 25x [96C 30sec, 65C 20min, 72C 30sec], 72C 2min, 4C hold. The amplification product was diluted 1:2,000 in water and 1 ul added to the Barcoding-PCR in a 10 uL reaction volume. The barcoded PCR products were pooled and the pools were purified using Ampure beads following the manufacturer's protocol, QCed on a Bioanalyzer DNA1000 chip (Agilent, Santa Clara, CA), and quantified using the Qubit dsDNA Broad Range kit (Thermo Fisher Scientific, Waltham, MA). Each pool was sequenced on a separate HiSeq 2500 Rapid run (Illumina, San Diego, CA) with 50 cycle paired end single index reads.

[0326] Tables 10-14 provide characteristics of the samples based on prior characterization. The number of samples per stage is shown in Table 10 based on histological analysis. Tables 11-14 are based on next generation sequencing analysis of tumor samples. The number of tumor samples with a CNV covering a at least 50% of the region is shown in Table 11. The number of tumor samples with a CNV covering at least 25% of the region is shown in Table 12. The number of regions with large abnormalities (at least 50% of the region) per patient is shown in Table 13. The number of regions with smaller abnormalities (at least 25% of the region) per patient is shown in Table 14.

Table 10. Patient Coverage in Tumor Samples

Stage	I	II	III	IV	All Malignant	Benign	Total
Num of Samples	11	10	11	8	40	40	80

Table 11. Summary Per Region (CNV covering 50%)

Chr	**Region**	**I**	**II**	**III**	**IV**	**Benign**
12	**1**	6	3	0	3	0
16	**3**	3	3	1	5	0
19	**4**	7	5	0	0	1
19	**5**	6	2	0	1	1
19	**6**	5	1	0	3	0
22	**7**	4	1	3	6	1
3	**8**	3	2	3	5	0
8	**9**	6	2	3	5	1

Chr	**Region**	**I**	**II**	**III**	**IV**	**Benign**
8	**10**	8	4	2	5	1
8	**11**	7	3	1	2	1
20	**12**	3	1	1	1	1
20	**13**	1	3	3	3	0
17	**14**	6	4	3	5	1
17	**15**	7	5	4	6	1

Table 12. Summary Per Region (CNV covering 25%)

Chr	**Region**	**I**	**II**	**III**	**IV**	**Benign**
12	**1**	7	3	1	4	1
16	**3**	3	5	2	6	0
19	**4**	7	5	1	1	1
19	**5**	6	4	0	2	1
19	**6**	7	1	0	3	0
22	**7**	6	4	3	7	1
3	**8**	3	2	3	5	1
8	**9**	7	2	3	7	1
8	**10**	8	4	2	5	1
8	**11**	8	3	1	5	1
20	**12**	4	3	2	1	1
20	**13**	4	3	3	4	1
17	**14**	6	4	4	5	1
17	**15**	7	5	4	7	2

Table 13. Summary of abnormalities per patient (abnormality at least 50%)

Stage	**>0**	**>1**	**>3**	**>5**	**All Samples**
I	9	9	9	5	11
II	5	5	5	5	10
III	5	5	4	3	11
IV	8	8	8	4	8
All					40
Malignant	27	27	26	17	
Benign	1	1	1	1	40

Table 14. Summary of abnormalities per patient (abnormality at least 25%)

Stage	**>0**	**>1**	**>3**	**>5**	**All Samples**
I	9	9	9	5	11
II	5	5	5	5	10

Stage	>0	>1	>3	>5	All Samples
III	5	5	4	3	11
IV	8	8	8	6	8
All					40
Malignant	27	27	26	19	
Benign	2	1	1	1	40

[0327] CNV was detected in 68% of tumor samples. We had two positives among benign samples, but one clearly seemed positive across all samples and one had a large duplication in the region in question.

Performance of ShapeIT

Summary per region

[0328] We calculated the errors made by ShapeIT to assess the effectiveness of our informatics haplotyping used in certain embodiments of methods herein. More specifically, among all the haplotype estimates made by ShapeIT between two consecutive heterozygous SNPs, we calculated the % of the SNPs where ShapeIT made an error. We also considered errors in SNP haplotypes where SNP calls were made with high confidence (>95% confidence) vs. low confidence errors (≤ 95% confidence). As shown in Table 15, ShapeIT errors were observed on all chromosomes tested, and errors were much higher in low confidence call samples.

Table 15. ShapeIt Error by Region

Chr	Region	Error %	High Conf Error %	Low Conf Error%
12	1	2.51%	0.61%	15.79%
16	3	1.58%	0.27%	11.66%
19	4	2.92%	0.37%	20.51%
19	5	1.07%	0.00%	11.96%
19	6	1.25%	0.23%	10.71%
22	7	2.03%	0.32%	16.51%
3	8	2.95%	0.45%	24.06%
8	9	3.08%	1.25%	18.68%
8	10	1.03%	0.59%	5.28%
8	11	1.20%	1.13%	1.99%
20	12	1.34%	0.84%	7.08%

Chr	Region	Error %	High Conf Error %	Low Conf Error%
20	**13**	1.47%	0.41%	10.24%
17	**14**	2.13%	0.47%	22.31%
17	**15**	3.60%	0.47%	29.92%

Summary per sample

[0329] Next, among the 9 cancer samples that were not contaminated, we compared the ShapeIT haplotypes with the haplotypes estimated from matched tumor samples. Haplotype estimation from tumor samples is believed to be accurate because the large allelic imbalance makes it relatively easy to determine haplotypes with high confidence. Table 16 provides ShapeIt results for each sample. Total error rate across all samples and all regions was 1.95%. The high confidence error rate was 0.60%, wherein the low confidence error rate was 14.25%.

Table 16: SharpeIt Error by Sample

Sample	Error %	HighConfError %	LowConfError%
DLS15-10446	3.02%	1.21%	21.30%
DLS14-23566	2.97%	0.80%	23.04%
DLS14-23548	1.76%	0.59%	16.17%
DLS14-23574	1.37%	0.30%	11.27%
DLS14-23570	2.02%	0.75%	15.66%
DLS15-10457	1.44%	0.32%	14.67%
DLS15-10447	1.53%	0.36%	8.88%
522	1.43%	0.43%	9.40%
528	2.10%	0.58%	16.44%

Performance of CNV Algorithm

[0330] We analyzed the data using a CNV algorithm with two main outputs: (1) Confidence and (2) Average allelic imbalance (AAI). When making a determination of copy number variability in a region, we used the confidence estimate (which is a function of the AAI estimate, number of SNPs, etc.). In plasma samples, CNVs were identified by a maximum likelihood algorithm that searched for plasma CNVs in regions where the tumor sample from the same individual also had CNVs using haplotype information deduced from the tumor sample. In the negative control samples, haplotype information was deduced from parental genotypes. The CNV detection algorithm modeled expected allelic frequencies across all allelic imbalance ratios at 0.025% intervals for three sets of hypotheses: (1) all cells are normal (no allelic imbalance), (2) some/all cells have a homolog 1 deletion or a homolog 2 amplification, or (3) some/all cells have a homolog 2 deletion or a homolog 1 amplification. The likelihood of each hypothesis was determined at each SNP using a Bayesian classifier based on expected and observed allele frequencies at all heterozygous SNPs, and then the j oint likelihood across multiple SNPs was calculated. Finally, the hypothesis with the maximum likelihood was selected. This algorithm also calculates the confidence of each CNV call by comparing the likelihoods of different hypotheses. A minimum confidence threshold of 99.9% was used in plasma samples from patients with cancer to minimize false-positive results. Further details regarding the analytical method used are provided in the section herein that discusses the Allelic_Analysis_Example.

[0331] We performed two sets of plasma runs, one with 1ml input DNA, the other one with 5ml input DNA.

1 ml input DNA runs:

[0332] Runs SQ1179-SQ1185 included a total of 28 samples (24 cancer samples and 4 normal model samples). Nine malignant, 2 benign, and 4 normal model samples (hence, 9 positives and 6 alleged negatives) were analyzed.

[0333] In tumor, a region was counted as positive for CNV if the CNV covered at least 25% of the region. We used a 95% confidence cutoff when calling a region positive in plasma. Based on that, the following table summarizes results.

Table 17: Results of CNV determinations (1 ml samples)

Sample	**Stage**	**Tumor CNVs**	**Plasma CNVs**
DLS15-10446	4	11	1
DLS14-23566	4	9	8
DLS14-23548	3	6	1
DLS14-23574	3	6	0
DLS14-23570	4	8	8
DLS15-10457	1	11	0
DLS15-10447	1	14	0
522	3	13	10
528	3	11	0
DLS14-23595	benign	0	0
DLS14-23531	benign	0	0
N020186-DNA	Normal	N/A	0
N020180-DNA	Normal	N/A	0
N020178-DNA	Normal	N/A	0
N029430-DNA	Normal	N/A	0

[0334] The maximum confidence indicating an abnormality in the negative samples was 86%, hence a 95% confidence threshold seems a conservative but reasonable choice. Further experiments and data may provide more evidence for decreasing the confidence cutoff for making a positive call (for example a confidence cutoff of 90% would have resulted in two plasma CNV calls in the samples where we had only one positive call with the 95% threshold, but it would not have changed the result for the samples with no positive calls). Accordingly, although in some embodiments a 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, and 99% cutoff are used, in illustrative embodiments a 90%, 91%, 92%, 93%, 94%, or 95% confidence cutoff is used.

[0335] The AAI estimate in the samples and regions where positive calls were made ranged from 1.39% to 14.91%. If the confidence cutoff were decreased to 90%, this range would have been 1.09% to 14.91%.

[0336] In certain embodiments, a no call range is defined as well. More specifically, a confidence range could be defined calls are not made on a region (e.g., in one embodiment <80% confidence is reported as low risk of CNV, 80% to 90% is reported as a no call, and >90% is reported as high risk). The specific ranges could be modified. For example, in one embodiment, less than 75% confidence in CNV is reported as low risk of CNV, 75% to 85% is reported as a no call, and greater than 85% is reported as high risk for CNV.

5 ml input DNA runs

[0337] Runs SQ1211 and SQ1212 included a total of 8 cancer samples. Three malignant and 1 benign sample were analyzed (i.e., 3 positives and 1 alleged negative).

[0338] Using similar methods and cutoffs as above to call positives in tumor and plasma, we obtained the results provided in Table 18.

Table 18: Results of CNV determinations (5 ml samples)

Sample	**Stage**	**Tumor CNVs**	**Plasma CNVs**
DLS15-10457	1	11	0
DLS15-10447	1	14	0
DLS14-23590	2	10	1
DLS14-23580	benign	0	0

[0339] The maximum confidence on an abnormality was 85% in the benign sample. Therefore, a 95% confidence cutoff for making a positive call again seems reasonable. However, in some embodiments, a 90% confidence cutoff is used.

Summary of all samples

[0340] The following Table 19, is a summary of positive call rate in the plasma summarized by cancer stage:

Table 19: Positive call rate in plasma by cancer stage

Stage	**Positive Calls**	**Total Samples**
I	0	2
II	1	1
III	2	4
IV	3	3
All Malignant	6	10
All Negative (Benign or Normal Sample)	0	7

[0341] Based on these results, the sensitive achieved was 60% and the specificity was 100%.

Conclusions

[0342] The selection of target sites for amplification, within haploblocks, yielded acceptable improved results for CNV detection in ctDNA. A high number of samples from malignant tumors did not exhibit any detectable abnormalities in the regions selected. This could be due to the biopsy or it could be due to the region selection.

[0343] ShapeIT performance for informatics haplotyping was acceptable and consistent with expectations. Furthermore, ShapeIT performance was consistent across patients.

[0344] The plasma CNV calling algorithm used in this embodiment did not detect CNVs in Stage I cancer samples, and did not detect all Stage 3 samples. It is possible that due to the biology of the ovarian tumors, the circulating free DNA amount in the plasma is not sufficient enough to catch certain CNVs. This is consistent with our observations related to SNVs. It is possible that further design improvements will provide sufficient sensitivity to detect CNVs in ctDNA in all Ovarian cancer patients. Nonetheless, the methods provided herein, which in illustrative examples as illustrated in Example 1, utilize pools of primers that target SNPs that are found within haploblocks and then utilize analytical methods with imperfect estimates of haplotypes, that take advantage of the fact that loci are selected that are within haploblocks, represent an important step in improved detection of CNVs in ctDNA in cancer.

<u>**Example 5**</u>

[0345] This example provides details regarding the identification of a panel of target chromosomal regions across eight driver genes, a primer pool for amplifying segments within such target chromosomal regions, which exhibit high somatic copy number variation (CNV) in lung cancer, wherein the primer pool is focused on primers that amplify SNPs within haploblocks, and analytical methods to assess copy number. The primer pool includes primer pairs (i.e. forward and reverse primers) for amplifying loci with strong linkage disequilibrium to other loci (i.e. loci within a set of haploblocks within target chromosomal regions known to exhibit CNV where a therapeutic has been identified), thereby useful for enrichment of target SNPs within haploblocks, for detecting CNV for a lung cancer therapy selection panel. The primer pairs are used to generate amplicons that can be analyzed, for example by high throughput sequencing. The primer pool was used to establish the feasibility of detecting lung cancer-relevant CNVs in plasma samples. The identified chromosomal regions in this design are focal CNVs and in fact, cover regions less than 2.5 megabases.

[0346] The Lung Cancer Therapy Selection Panel analyzed in this Example is a RUO liquid biopsy test targeted towards patients with a known diagnosis of lung cancer. It focuses on multiple types of lung cancer alterations that impact therapy decisions and detects single nucleotide variations (SNVs), copy number variations (CNVs), and gene fusions. The panel is intended to be used on plasma cfDNA samples.

[0347] In particular, this example illustrates the analytical performance of the focal CNV (fCNV) component of this test. Focal CNVs in this example are generally covering short regions (<2.5 Mb). The current version of the Lung Cancer Therapy Selection Panel aims to detect fCNVs surrounding eight targeted genes including BRAF, EGFR, ERBB2, FGFR1, KRAS, MET, MYC, and PIK3CA.

[0348] This example also provides detailed copy number determinations obtained by analyzing the samples using the quantitative, non-allelic FODDOR method, and illustrates the complementary nature of a quantitative, non-allelic method like the FODDOR method with the allele-based haploblock method.

[0349] The FODDOR algorithm can be used for classifying a sample as either positive or negative. This is done by checking if all the regions of interest in the sample have the same copy number or if they have different copy numbers.

[0350] In addition to classifying a sample as positive or negative, FODDOR can also estimate the virtual tumor fraction (VTF) of the region with maximum abnormality. VTF of a region is defined as the tumor cell fraction of a tumor with copy number equal to 3 in that region that is required to generate the copy number observed in that region. That is, suppose that a hypothetical tumor has a copy number equal to 4 in just one abnormal region and suppose that tumor's cell fraction in the corresponding plasma is 0.05. The VTF of this region is the tumor cell fraction that is required to generate the equivalent excess of this region assuming that the region's average copy number is 3. The conversion from VTF to TCF is given by: VTF = (N-

2)xTCF, where N is the average copy number of that region in the tumor. We estimate the VTF by estimating the excess of a region compared to the average of all the other regions.

[0351] Using the above two features of FODDOR, we also designed an estimator which, subject to certain conditions, can make calls on and estimate the individual region copy numbers. This is done by iteratively running FODDOR to pick out one abnormal region per iteration until FODDOR identifies a subset of regions that all have the same copy number. In case FODDOR cannot identify a subset of at least two regions with same copy number, the sample is no-called. More information about the FODDOR method is provided in a separate section in this specification.

[0352] The fCNV panel includes genes with recurrent fCNVs that are demonstrated to have clinical utility in the treatment of patients with lung cancer. This utility is based on meeting at least one of the following criteria: (1) Credentialed per NCCN guidelines or FDA-labeling for selection of an approved treatment target in a lung cancer; (2) Credentialed per NCCN guidelines or FDA-labeling for selection of an approved treatment target in any malignancy but robust clinical data are lacking demonstrating efficacy in lung cancer (i.e. "Off-Label"); (3) The mutation is an eligibility criteria for an ongoing clinical trial (per ClinicalTrials.gov).

[0353] The eight target genes that are all amplified oncogenes that are targets of existing therapeutic agents (FDA-approved use or off-label use), or therapeutic agents in development (clinical or pre-clinical), were identified (See Table 20). Of the 8 genes, MET amplification is credentialed (category 2A) in NCCN Non-Small Cell Lung Clinical Guidelines (version 6.2015) as an emerging target for Crizotinib treatment (Ou, 2006). For these genes, target regions that are known to be amplified in lung cancer are shown in Table 20 along with a therapeutic targeted to the gene with the CNV.

Table 20. Target lung cancer genes, chromosomal regions, and justification of therapeutic utility

((^a^FDA-Approved/^b^Off-Label/^c^Clinical Trial/^d^Preclinical)			
Gene Name	**Gene Coordinates (hg19)**	length [kb]	**Indicated Targeted Therapeutic**
BRAF	chr7:140433813-140624564	191	vemurafenib[b]; dabrafenib[b] (approved for gene mutations)
EGFR	chr7:55086725-55275031	188	cetuximab[a]; erlotinib[a]; gefitinib[a]; afatinib[a]; panitumumab[b]; vandetanib[b]; lapatinib[b]
ERBB2	chr17:37856231-37884915	29	afatinib[a]; ado-trastuzumab emtansine[b]; pertuzumab[b]; trastuzumab[b]; lapatinib[b]
FGFR1	chr8:38268656-38325363	57	ponatinib[b]
KRAS	chr12:25358180-25403854	46	Mekinist[a] Selumetinib[c] (for gene mutations)

((aFDA-Approved/bOff-Label/cClinical Trial/dPreclinical)			
Gene Name	**Gene Coordinates (hg19)**	length [kb]	**Indicated Targeted Therapeutic**
MET	chr7:116312459-116438440	126	crizotinib[a]; cabozantinib[a]
MYC	chr8:128748315-128753680	5	gefitinib (high copy number may confer increased EGFR tyrosine kinase sensitivity)
PIK3CA	chr3:178866311-178952497	86	Dactolisib[c] Buparlisib[c] (for gene mutations)

[0354] Target chromosomal regions of each gene of interest were identified based on the following considerations: There are three main papers that studied lung cancer by analyzing large number of samples, each on a different subtype of Lung cancer:

TCGA 2012 - 178 SQCC (Lung Squamous Cell Carcinoma) samples (Nature 489:519-25 (2012). doi:10.1038/nature11404);

TCGA 2014 - 230 ADC (Lung Adenocarcinoma) samples (Nature 511:543-50 (2014). doi:10.1038/nature13385); and

George et al. 2015 - 110 SCLC (Small Cell Lung Cancer) samples (Nature 524:47-53 (2015). doi: 10.1038/nature14664).

[0355] Table 21 presents regions identified by three main lung cancer studies with statistically significant focal copy number alteration (q-value < 0.05) for the eight target genes reported in these studies.

Table 21. Chromosomal regions with focal copy number alterations

Subtype	**Gene**	**Chr**	**Start (hg19)**	**End (hg19)**	**Length (Mb)**	**q-value**	**CNV type**
ADC	KRAS	12	25402469	26433911	1.03	1.330E-05	Amp
ADC	EGFR	7	54535672	55737616	1.20	1.520E-05	Amp
ADC	MET	7	116283302	116449049	0.17	5.913E-04	Amp
ADC	MET*	7	115368861	117051327	1.68	2.483E-04	Amp
ADC	ERBB2	17	37804811	38011853	0.21	1.902E-02	Amp
SQCC	FGFR1	8	38170522	38286018	0.12	1.19E-30	Amp

Subtype	Gene	Chr	Start (hg19)	End (hg19)	Length (Mb)	q-value	CNV type
SQCC	MYC	8	128202879	128788635	0.59	6.79E-10	Amp
SQCC	EGFR	7	54642932	55858372	1.22	8.85E-07	Amp
SCLC	PIK3CA	3	178430118	186909171	8.48	2.44E-10	Amp

[0356] We also investigated the dataset of COSMIC ASCAT CNV Events to inspect recurrent CNV regions covering the target genes. Cosmic uses ICGC CNV profiles where available, otherwise Cosmic reanalyzed TCGA with ASCAT. ASCAT accounts for normal cell admixture and tumor aneuploidy in CNV estimation using B-allele frequencies. Note that COSMIC is more conservative with respect to CNV calls. Table 22 provides the results from the COSMIC analysis. The CNVaffected regions in each gene vary among patients and they are longer than the coding region of the gene. As can be seen in the table below, the average length of CNV region per each gene is from 0.5Mb to 33Mb. We also observe that the majority of CNVs in the target genes are high level amplification (with median copy number (50th percentile ≥ 9) (Table 22).

Table 22. Recurrent CNV regions from analysis of COSMIC lung cancer data

		copy number			Major to Minor Haplotype ratio			length [Mb]		
Gene Name	# Samples from Cosmic	5th PCTL	50th PCTL	90the PCTL	5th PCTL	50th PCTL	90the PCTL	5th PCTL	50th PCTL	90the PCT L
BRAF	19	2	9	14	1.3	2.0	9.2	1.10	33.04	50.50
EGFR	85	5	12	36	1.0	5.0	22.0	0.46	3.09	49.15
ERBB2	18	7	16	62	1.4	5.0	18.5	0.20	0.50	10.33
FGFR1	110	1	10	27	1.3	5.8	16.7	0.37	1.76	15.31
KRAS	61	5	10	19	1.2	4.0	13.0	0.31	5.37	27.12
MET	37	5	10	37	1.2	4.0	12.0	0.38	7.53	49.95
MYC	107	5	10	24	1.0	4.5	13.0	0.12	2.01	30.65
PIK3CA	182	5	9	18	1.0	4.0	9.3	1.46	11.96	40.85

[0357] We applied additional processing to these reported regions to determine target regions for our panel:

KRAS - reported statistically significant region by TCGA for subtype ADC has been chosen as the target region.

MYC - reported statistically significant region by TCGA for subtype SQCC has been chosen as the target region.

EGFR - the overlap between the two statistically significant regions identified by TCGA for subtypes ADC and SQCC has been considered as the target region.

MET - the region reported in Table 21 is small and therefore not feasible for design. However, the same study identified a larger region (1.68Mb) including MET with statistically significant CNV for a sub-group of patients (n=87). We decided to choose that region as the target region.

[0358] It is noteworthy that TCGA considered these genes as "drivers": KRAS, EGFR, ERBB2, BRAF, MET, ALK fusion genes, RET fusion genes, ROS1 fusion genes, HRAS, NRAS, and MAP2K1.

[0359] Next, mutations were filtered to include only those with either evidence of recurrence within the COSMIC databases (>3 independent mutations at the same site) or evidence of functional impact (e.g. MAP2K1 p.C121S4 and MET exon 14 deletions 5, 6).

[0360] After mutation filtering, we considered any sample having a mutation in one of the above listed genes listed as belonging to the "oncogene-positive" group (n = 143). Samples lacking any of the mutations were considered "oncogene-negative" (n = 87).

[0361] ERBB2 and FGFR1 - the region reported in Table 21 is small and therefore not feasible for design. COSMIC data shows the CNV regions are quite variable for these genes, which makes it hard to identify the most common CNV region. Therefore, we decided to target a window of 1.5Mb around the gene.

[0362] PIK3CA - reported region for SCLC is quite large. We chose a common CNV region among 80% of samples (141 out of 177) in COSMIC data. Note that we filtered for non TCGA samples or loss CNVs in Cosmic data.

[0363] BRAF - there is no statistically significant CNV region reported in the literature. We chose a common CNV region among 80% of samples (13 out of 16) in COSMIC data. Note that we filtered for non TCGA samples or loss CNVs in Cosmic data.

[0364] Based on the above considerations, the chromosomal regions that were selected as target chromosomal regions are shown in Table 23.

Table 23. Selected target chromosomal regions

Gene	**Chr**	**Start (hg19)**	**End (hg19)**	**Length (Mb)**	**Total SNPs MAF >=.1**
KRAS	12	25402469	26433911	1.03	2184
EGFR	7	54642932	55737616	1.09	2377
MET	7	115368861	117051327	1.68	2267
FGFR1	8	37500000	39000000	1.50	1740

Gene	Chr	Start (hg19)	End (hg19)	Length (Mb)	Total SNPs MAF >=.1
PIK3CA	3	178431895	179540177	1.11	1820
MYC	8	128202879	128788635	0.59	1432
BRAF	7	138448946	140783654	2.33	3631
ERBB2	17	37000000	38500000	1.50	1844

SNP Loci And Primer Design Requirements

[0365] The following pool design requirements were specified:

Target the top regions in ovarian cancer such that at least 80% of patients reported in TCGA are covered;

SNPs should be part of relatively large haplotype blocks such that the informatics phasing error rate is less than 5% on average for each region of interest;

SNPs covering specific cancer-related genes in regions of interest should be given high priority;

At least 1,000 SNPs should be identified per target chromosome region;

All primer designs compatible with mmPCR in one pool, meaning all dual extensible interactions in one pool, should have deltaG higher than -4 kcal/mol;

The SNP target loci should be located in the first 50 bases of amplicon;

The SNP loci allele determination should be compatible with HiSeq 2500 50bp singleend sequencing (note that not all assays necessarily satisfy the Nextseq 75bp paired-end requirements);

[0366] The following were the main primer design requirements:

One pair of left and right primers per target SNP;

Optimal Tm 56C, allowed range [53C-59C];

Amplicon length 50-75 bp;

GCcontent 30-70%;

Maximum GC clamp 4;

Pool Design;

[0367] The design process consisted of these main steps:

Select candidate target SNPs for each region of interest;

Attempt to design up to five sets of right and left specific primers for each candidate target SNP;

Identify known haplotype blocks for SNPs with a design;

Select compatible designs to form the primer pool(s);

Candidate SNPs Selection

[0368] For each region of interest we chose candidate SNPs satisfying following criteria:

The SNP must be present in both dbSNP Common 138 and the 1000 Genomes project (the phase 1 version 3 variant calls released April 30, 2012) variant call data set;

The SNP minor allele frequency from the 1000 Genomes project must be at least 10%;

The SNP location must be within one of the corresponding breakpoints in Table 20.

Primer Design

[0369] The primers were designed using Primer3 release 2.3.6 and the RunPrimer3 Java program using the design parameters in Table 24. For each candidate SNP Primer3 was used to design left and right primers (two-sided) with amplicon length within a range of 50 to 75 bp and melting temperature between 53-59°C optimized at 56°C. Primer3 was configured to use the SantaLucia salt correction and melting temperature formulae. Primer locations were restricted to be at least 2bp away from any SNP which is present either in dbSNP Common 138, or in the 1000 Genomes project with minor allele frequency greater than 1%. Up to five designs can be generated per target. Since previously we did not identify an issue for test primers with 4 GC clamp, we decided to limit the GC clamp to 4.

Table 24. Primer design parameters

Name	Value	
target_padding	2	Primers should end at least 2 bases away from the target loci
PRIMER_MAX_SIZE	30	
PRIMER_OPT_SIZE	24	
PRIMER_MIN_SIZE	18	
PRIMER_WT_SIZE_LT	0.5	
PRIMER_WT_SIZE_GT	0.5	Penalty for primer longer than optimal
PRIMER_PAIR_WT_PRODUCT_SIZE_LT	0	
PRIMER_PAIR_WT_PRODUCT_SIZE_GT	1	Penalty for amplicon longer than optimal
PRIMER_MAX_TM	59	
PRIMER_OPT_TM	56	
PRIMER_MIN_TM	53	
PRIMER_WT_TM_LT	1	Penalty for TM lower than optimal
PRIMER_WT_TM_GT	1	Penalty for TM higher than optimal
PRIMER_MAX_GC	70	
PRIMER_OPT_GC_PERCENT	50	
PRIMER_MIN_GC	30	
PRIMER_WT_GC_PERCENT_LT	1	
PRIMER_WT_GC_PERCENT_GT	1	
PRIMER_MAX_END_GC	4	
PRIMER_MAX_POLY_X	5	
PRIMER_INTERNAL_MAX_POLY_X	5	
PRIMER_SALT_CORRECTIONS	1	
PRIMER_SALT_DIVALENT	0	
PRIMER_DNTP_CONC	0	
PRIMER_THERMODYNAMIC_OLIGO_ALIGNMENT	1	
PRIMER_THERMODYNAMIC_TEMPLATE_ALIGNMENT	1	
PRIMER_MISPRIMING_LIBRARY	Human	The mispriming library containing microsatellites downloadable from

Name	Value	
		Primer3 website.
PRIMER_LIB_AMBIGUITY_CODES_CONSENSUS	1	

[0370] We skipped the filtering for the probable mispriming. We found that mispriming filtering was too stringent and it over-filtered designed primers. Finally, if a SNP target has multiple designs we chose the design with the shortest amplicon length.

Haplotype Block Identification

[0371] We used a program called plink (v1.90b3p 64-bit (10 Oct 2014)) to identify haplotype blocks for our regions of interest based on the definition provided herein. The program has been run for each region separately on the set of SNPs with proper designs produced in the previous step. The 1000 genomes project haplotypes release on 2013-09 was used at the reference panel. The release contains haplotypes on 1092 samples (#haplotypes = 2184) for 36.8 million SNPs.

Pooling

[0372] The final step of the design process was to choose a subset of the candidate SNPs with designs that could be combined into a single multiplex primer pool. To be able to pool the set of designed primers we needed to minimize the possibility of primer-dimer formation. The tendency of two primers to bind to each other can be estimated by the Gibbs free energy and/or the melting temperature of their most stable interaction.

[0373] For every pair of primers in the design set we calculated the Gibbs free energy (deltaG) and the corresponding melting temperature (Tm) for three types of interactions including the strongest dual extensible, the extensible, as well as any. An extensible interaction is defined as one with at least three base matches at the 3'end of the primer. All calculations were based on a thermodynamic approach using the following design parameters:

temperature = 56 C;

primer_concentration = 50 nM;

salt_concentration = 50 mM;

forward_tag=ACACGACGCTCTTCCGATCT;

reverse_tag=AGACGTGTGCTCTTCCGATCT;

[0374] The interaction score for each pair of primers was set to max{deltaG2, 90%*deltaG12, 65%*deltaG012}. Based on prior experience we believe primers with interaction score weaker than -4 kcal/mol are less likely to create primer-dimers, and thus can be in a multiplex primer pool.

[0375] We ran a pooling algorithm that analyzed primer dimers with the above considerations to select an optimized set of designs with no high-scoring interactions (<-4 kcal/mol). The algorithm is a heuristic method that attempts to choose a required compatible number of SNPs from large halplotype blocks. Based on simulation results in Example 3 SNPs in haplotype blocks smaller than 10 were less likely to contribute to the CNV detection algorithm. Therefore, we decided to ignore any block smaller than 10. The utility score of a target includes the following weighted factors: number of patients having a CNV spanning over the SNP location (w=.5); the haplotype block size that a SNP belongs to (w=.2), target SNP minor allele frequency (w=.3), observed heterozygosity rate (w=. 1), presence in HapMap (w=. 1), transversion mutation (w=. 1), amplicon GC-content (w=.1) and amplicon length (w=.1). The pooling algorithm first builds a conflict graph, where assays are nodes and the edge between two nodes represents a high score interaction between the primers of corresponding assays. Then it tries to find the Maximal Independent Set by iteratively removing the highest degree node at each step. In case there are several nodes with highest degree, the one with the lowest utility score is removed.

General Methodology

[0376] In this study we analyzed cell-line derived cfDNA titrations and plasma from healthy individuals. Unless indicated otherwise, sample preparation and sequencing analysis was performed as set out in Example 4. Briefly, samples were made into libraries by ligation of adapters followed by PCR to amplify the available cfDNA. The selected SNPs in the target genetic regions were then amplified by massively multiplexed PCR. The amplification protocol for multiplex PCR was as follows: 95C 15 min, 17x[95C 30sec, 62.5C 15min, 72C 5min], 72C 2min, 4C hold.

[0377] The resultant amplicon pool was sequenced using next generation sequencing and the resulting data was analyzed to determine the presence of fCNVs in the target genes that are listed in Table 23.

Scope

[0378] Artificial cfDNA samples were generated with known relative copy number changes that ranged from above to below the expected limit of detection of our method, resulting in <1% to

>40% expected average allelic imbalance (AAI). These known positive samples were then used to assess the sensitivity of our technology.

[0379] Specificity was tested using both negative artificial cfDNA and cfDNA extracted from standard plasma samples from healthy individuals.

[0380] Exemplary abbreviations used specifically in this Example:

AAI
Average Allelic Imbalance;
fCNV
focal CNV;
FODDOR
Focal CNV Detection using Depth of Read;
NAT
normal adjacent tissue;
NCCN
National Comprehensive Cancer Network;
NGS
Next Generation Sequencing;
NIPT
Non-invasive prenatal testing
Plasmart
Artificially created plasma sample;
SNV
Single Nucleotide Variation;
TCF
Tumor Cell Fraction;
VTF
Virtual Tumor Fraction.

Samples Description

[0381] Three pairs of matching (one pair per individual) CNV-affected tumor and non-affected wild type cell lines were purchased from ATCC and cultured according to ATCC recommendations.

[0382] The presence of CNVs was confirmed using Oncoscan and NGS data. More specifically, the regions shown in Table 25 were found to be good candidates in each cell line (i.e., they had obvious copy number differences between the homologs).

Table 25 Samples used in titration experiment

Matched Cell Line Pairs	Regions
Cell Line Pair A	EGFR, ERBB2, FGFR1, KRAS, MET, PIK3CA
Cell Line Pair B	BRAF, ERBB2, FGFR1, MET
Cell Line Pair C	BRAF, ERBB2, FGFR1, KRAS, MYC, PIK3CA

[0383] These titrations simulate the stated tumor cell fractions (TCF) of 1%, 2%, 3%, 5%, 7%, 10%, and 20%. A sample with 1% TCF refers to a sample containing DNA from 1 tumor cell per 99 wild type cells.

[0384] These synthetic samples simulate cfDNA extracted from plasma of cancer patients with known CNVs and were used to determine the limit of detection based on known TCF. Note that the level of abnormality is unknown in real cancer plasmas, hence they cannot be used to determine the limit of detection.

[0385] Negative control libraries were generated from both mononucleosomal DNA from wild type cell lines and from cfDNA extracted from standard plasma samples from healthy individuals.

Matching tumor and normal cell lines

[0386] Pairs of matching tumor and normal cell lines were generated from the same individual cancer patient and were purchased from ATCC. Cell lines were not selected for tumor origin but data availability in public databases that indicated CNVs affecting the coding region of assay panel covered target genes. A list of the selected cell lines and additional information such as tissue origination and cancer stage are shown in Table 26.

Table 26. Tumor cell lines used in this study Gazdar et. al. 1998)

cell line	**tissue**	**primary stage**	**cell line characteristics**	**patient**
HCC1954	mammary gland; breast/duct; epithelial	IIA, grade 3 invasive ductal carcinoma with no lymph node metastases	poorly differentiated cell line initiated on October 30, 1995; it took about 4 months to establish	61 years adult, East Indian, Female
HCC2218	mammary gland; breast/duct; epithelial	TNM stage IIIA, grade 3,primary invasive ductal carcinoma with metastases in 42/43 lymph nodes	poorly differentiated cell line initiated on April 10, 1996, and took 6 months to establish	38 years, Caucasian, White, Female
HCC38	mammary gland; breast/duct;	TNM stage IIB, grade 3, primary ductal carcinoma	initiated on 4/27/92 and took 32 months to establish	50 years, Caucasian, White,

cell line	tissue	primary stage	cell line characteristics	patient
	epithelial			Female

Normal reference cell lines

[0387] Normal reference cell lines are generated from leukocytes of the cancer patient by EBV-transformation.

Tumor cell lines

[0388] Matching tumor cell lines are made from various kinds of tumor tissues or metastases by months of repeated subclonation. This process can cause subclonal CNV and SNV occurrence within a cell line during cultivation and causes CNVs of larger genome regions than commonly seen in true tumor biopsies. However, genome rearrangements in DNA samples extracted from the same culture have CNVs that remain constant throughout experiments conducted with those samples.

Artificial cfDNA preparation

[0389] We used the MNase-based shearing of cell line DNA into mononucleosomal DNA fragments to simulate cfDNA. Mononucleosomal DNA (150 bp fragments) from each of these CNV-affected and non-affected cell line pairs was purified and mixed to generate a range of known CNV titrations.

[0390] DNA samples were characterized with Oncoscan to establish the exact CN for each genome region. Tumor and normal DNA were titrated over a range of tumor fractions to create artificial samples. These have a known CNV copy number and tumor fraction for each CNV in each sample.

DNA yield consideration

[0391] Reference cell lines grown in suspension at high cell counts and high yield for mononucleosomal DNA were used as artificial cfDNA. In contrast to this, adherent growing tumor cell lines have lower cell counts per culture and MNase-treatment yields much less mononucleosomal DNA.

Considerations on compatibility of cell lines with bias model

[0392] Artificial samples prepared from cell lines have previously shown performance inconsistent with patient plasma, suspected to be due to differences in resulting data characteristics. A simple method to measure similarity between artificial samples and a set of reference data such as real plasma is to compare the distribution of reads over the individual targets. This can be computed as a correlation coefficient between average per-target amplification rates, calculated between a set of artificial samples and a set of reference samples. Table 27 shows the correlation coefficients calculated for various data sets compared to their corresponding references.

Table 27. Amplification correlation coefficients for various data sets measured against corresponding reference data

Data set	**Amplification correlation coefficient against reference data**
Microdeletions validation study plasmart	0.96
Panorama V3 feasibility study plasmart	0.88
Panorama commercial data affected by poor quality extraction reagents	0.87
Focal CNV cell line titration study	0.87

Samples

[0393] CNVs including "focal" CNVs are larger than hundreds of kb in length which is too large to use synthetic DNA to generate artificial DNA samples with known CNVs, similar to what was used for SNV-spikes. Additionally, the AAI-approach requires normal and CNV-affected samples to have the same SNP-pattern and to be derived from the same donor. Two kinds of test samples used for this study fulfill these requirements: cell lines and lung cancer patient samples.

Tissue DNA preparation for reference experiments

[0394] Four FFPE- and 3 8 FF matching sample sets of lung cancer patients of various carcinoma types and stages purchased from CRO were included in this experiment as shown in Table 28. Tumor- and normal tissue DNA and plasma cfDNA were extracted and used for subsequent analysis.

Table 28. Overview of patient samples.

Histological diagnosis	sample count		Stage	sample count		Highest Stage	sample count
adenocarcinoma	15		I	9		IA	9
small cell carcinoma	2		II	18		IB	21
squamous cell carcinoma	20		III	11		IIA	0
bronchioloalveolar adenocarcinoma	3		IV	0		IIB	7
adesquamous carcinoma	2		n/a	4		IIIA	4
total	42		total	42		IIIB	0
						IV	1
						total	42

Sample preparation

[0395] Mononucleosomal DNA from cell lines was prepared according to the protocol described in Wapner et al. 2014 and mixed.

Library preparation

[0396] The titration and real cfDNA samples were converted into libraries using the Natera library preparation kit. Libraries were prepared from the cell line MNased DNA samples, cell line titrations, and patient plasma cfDNA. The cell line derived sample and titration libraries contained 10k haploid genome copies (~33 ng) of DNA input material. Due to the large variance in total cfDNA available per patient, one library per patient was prepared with 40 ul cfDNA. All libraries were made with 15 cycles of library amplification and were purified using AMPure (Beckman Coulter, Brea, CA).

Multiplex PCR

[0397] The multiplex PCR protocol as disclosed in Example 4 above, was performed on each library using the Lung fCNV Primer Pool except that a 62.5°C annealing temperature was used. Accordingly, the cycling conditions for the multiplex PCR was as follows:
95C 15 min, 17x[95C 30sec, 62.5C 15min, 72c 5min], 72C 2min, 4C hold Multiplex reactions using titration and plasma cfDNA libraries contained 6.7 ul of purified library input while the pure cell line reactions used 3 ul of purified library. Each reaction was done in triplicate. Each

reaction contained approximately 15k haploid genome copies (50 ng).

Barcoding PCR, Pooling and Sequencing

[0398] Each OneSTAR PCR reaction was barcoded. To fit the needs of AAI analysis the titration and plasma cfDNA reactions were pooled with 16 samples per pool to maintain an average DOR/assay > 4,000. The titration and plasma cfDNA reactions were pooled into two additional and separate FODDOR pools containing 240 samples each. This creates an average DOR/assay of 290. Barcoded reactions of cancer-free patient plasmas used in the Bias Model experiment were included in the two FODDOR pools to ensure a final reaction count of 240 samples per pool. The patient tumor and normal tissue barcoded reactions were pooled with the pure cell line reactions and had an approximate DOR/assay of 615.

fCNV workflow

[0399] Library products were subj ected to the fCNV workflow and products were barcoded and pooled. The pools were quantified and sequenced. The sequence data was analyzed to determine sensitivity and specificity.

Defining true CNV status for use as reference

[0400] Two external methods were used to establish CNV-truth for cell lines and tissues, OneSTAR Truth and Oncoscan. In addition, we also sequenced the tumor cell-line as a genomic sample and as a library.

OneSTAR Truth

[0401] OneSTAR PCR with the lungTSP fCNV panel was used to measure AAI in DNA samples from tumor/wild type cell lines and tumor/normal reference tissues. This method provides AAI but not an absolute CN per CNV.

Oncoscan

[0402] Oncoscan uses a different, much larger set of SNP-probes than the lungTSP fCNV panel to estimate genome-wide CNs (Table 29). The CN estimate is based on both allele frequency and probe intensity.

Table 29. The number of SNP-probes present within each gene region in NGS and Oncoscan.

Gene	**Chr**	**StartPos**	**EndPos**	**SNPs in NGS**	**SNPs in Oncoscan**
BRAF	7	138449419	140782039	836	304
EGFR	7	54646322	55737172	611	272
ERBB2	17	37000013	38496752	497	269
FGFR1	8	37501860	38993605	495	149
KRAS	12	25404604	26430452	493	78
MET	7	115376764	117048082	539	286
MYC	8	128203857	128788247	426	157
PIK3CA	3	178435382	179540177	430	120

[0403] Oncoscan was used to establish CNVs and CNs for the genomes of the tumor cell lines used for titrations, and to make predictions about AAI in cfDNA titration samples. Oncoscan was not used for patient tumor samples.

COSMIC

[0404] The Cosmic database was used to initially choose tumor cell lines with CNVs in the assay covered regions. The DNA preparations of these cell lines were then validated by Oncoscan.

[0405] In several cases the target gene coding region fell into a gap between CNVs reported in Cosmic, leading to an absence of a CN-call in Cosmic. However, Oncoscan data of the same cell line for the same region shows a continuous CNV. This is probably caused by Cosmic only annotating the highest CN of a region while slightly less affected regions are not annotated as CNV-affected, leading to many false negative calls. Discrepancies between Cosmic and Oncoscan CN-calls for the genes and cell lines used in this study are listed in Table 30.

[0406] Cosmic and Oncoscan agree on the locations of CN-transitions in these regions, which can be interpreted as a) the cell line genome is stable enough to allow reproduction of results between the experiment represented in Cosmic data and our DNA prep and

b) the reported gaps in Cosmic are most likely misrepresentations and the not reported CNs are false negatives.

Table 30. Comparison of COSMIC and OncoScan CN-calls ("Cosmic CN" / "Oncoscan CN")

COSMIC/ Oncoscan CN Call	BRAF	PIK3CA	MYC	MET	KRAS	FGFR1	ERBB2	EGFR
HCC1954	2/2.3	2/3	2/10	2/2.6	2/2	2/2	14/69	2/3
HCC2218	2/4	2/2	2/6	2/4	2/2	2/2	14/23	2/2

COSMIC/ Oncoscan CN Call	BRAF	PIK3CA	MYC	MET	KRAS	FGFR1	ERBB2	EGFR
HCC38	2/2.3	2/2.6	2/3.3	2/2	2/2	2/2	2/2	2/2

Truth used for final analysis

[0407] As explained previously, we estimated the true copy number and AAI of each of the regions in the three cell-lines using several different techniques. The results were not completely concordant, but were merged into a final "best estimate" truth that we used for analyzing the performance of the algorithms. Table 31 lists this "best estimate" truth. All of the regions are considered affected by a CNV except as described below the table.

Table 31. Best estimate of true copy number used as reference for performance analysis

	BRAF	EGFR	ERBB2	FGFR1	KRAS	MET	MYC	PIK3CA
HCC38	2.06	2[1]	2	2.33	2.275	2[2]	3.33	2.67
HCC1954	3[3]	3	37.38	0	0	3	6.59[4]	2.85
HCC2218	2.67	2[5]	6.32	1	2[6]	3.00	4[7]	28

[1] EGFR (not >99% confident about >0% AAI in the TCF=20% sample despite clear abnormality in tumor).

[2] MET (0% AAI in the TCF=20% sample despite clear abnormality in tumor).

[3] Excluded region. BRAF (0% AAI in the TCF=20% sample despite clear abnormality in tumor.

[4] Excluded region. MYC (balanced duplication covering most of the region, with a small unbalanced duplication).

[5] EGFR (CN=2 in tumor).

[6] KRAS (CN=2 in tumor).

[7] MYC (balanced duplication).

[8] PIK3CA (CN=2 in tumor).

Average allelic imbalance algorithm

[0408] An improved version of the CNV calling algorithm described at Kirkizlar et al. 2015 (Kirkizlar et al., Translational Oncology, 8 407-416) was used to make the fCNV calls. The algorithm uses haplotype information estimated through informatic methods rather than the perfect haplotype information obtained through tumor samples. Note that haplotype information predicts which alleles are present on a single chromosome homolog and would therefore be present with the same homolog copy number.

[0409] Briefly, the algorithm computes an average AAI value that fits the data best at each region together with the corresponding confidence. We use the AAI and confidence values together to make the final call.

[0410] More specifically, we analyzed the data using a CNV algorithm with two main outputs: (1) Confidence and (2) Average allelic imbalance (AAI). When making a determination of copy number variability in a region, we used the confidence estimate (which is a function of the AAI estimate, number of SNPs, etc.). In plasma samples, CNVs were identified by a maximum likelihood algorithm that searched for plasma CNVs in regions where the tumor sample from the same individual also had CNVs using haplotype information deduced from the tumor sample. In the negative control samples, haplotype information was deduced from parental genotypes. The CNV detection algorithm modeled expected allelic frequencies across all allelic imbalance ratios at 0.025% intervals for three sets of hypotheses: (1) all cells are normal (no allelic imbalance), (2) some/all cells have a homolog 1 deletion or a homolog 2 amplification, or (3) some/all cells have a homolog 2 deletion or a homolog 1 amplification. The likelihood of each hypothesis was determined at each SNP using a Bayesian classifier based on expected and observed allele frequencies at all heterozygous SNPs, and then the joint likelihood across multiple SNPs was calculated. Finally, the hypothesis with the maximum likelihood was selected. This algorithm also calculates the confidence of each CNV call by comparing the likelihoods of different hypotheses. A minimum confidence threshold of 99.9% was used in plasma samples from patients with cancer to minimize false-positive results. Further details regarding the analytical method used in this Example are provided in the analytical method called the Allelic_Analysis_Example discussed herein.

[0411] AAI can be interpreted as the average difference between the copy numbers of the homologs, and is analogous to the variant allele frequency in SNV detection. The reason behind using AAI as the main performance measure is due to the fact that the TCF can be ambiguous for the regions with multiple abnormalities. In order to relate AAI to TCF, one could assume that a region has a constant copy number, for example, one extra copy throughout the region, and then compute the corresponding TCF from the observed AAI. Table 32 below shows the relationship between AAI and TCF under the assumption that one homolog always has one copy and the second homolog is amplified.

Table 32. AAI as a function of TCF and tumor copy number

TCF	CN = 3	CN = 4	CN = 5	CN = 6
1%	0.50%	0.99%	1.48%	1.96%
2%	0.99%	1.96%	2.91%	3.85%
3%	1.48%	2.91%	4.31%	5.66%
5%	2.44%	4.76%	6.98%	9.09%
7%	3.38%	6.54%	9.50%	12.28%

TCF	CN = 3	CN = 4	CN = 5	CN = 6
10%	4.76%	9.09%	13.04%	16.67%
15%	6.98%	13.04%	18.37%	23.08%
20%	9.09%	16.67%	23.08%	28.57%

[0412] The Table presented in FIG. 8 provides the AAI estimate as a function of TCF values and total copy number of the tumor cells (assuming an unbalanced duplication where one homolog has one copy). Note that due to the mosaic nature of the cell lines and complex duplication patterns, FIG. 8 only provides an approximation to our observed AAI.

[0413] As for AAI method, we use the average AAI estimate at 20% TCF as the truth (Table 33).

Table 33. Observed AAI at 20% TCF

AAI%	BRAF	EGFR	ERBB2	FGFR1	KRAS	MET	MYC	PIK3CA
HCC38	0	3.73	81.82	6.34	10.37	3.64	1.08	7.06
HCC1954	10.50	1.31	16.87	4.99	1.12	19.29	0.86	0
HCC2218	7.18	1.72	19.76	6.49	2.7	0	2.68	17.19

DNA Extraction

[0414] Genomic DNA from tumor and normal cell lines were extracted and enzymatically fragmented into "MNased DNA."

[0415] cfDNA was extracted from each of the 42 patient plasma samples using the QIAamp Circulating Nucleic Acid kit (Qiagen, Hilden, Germany) and was eluted in 50 ul of DNA Suspension Buffer. DNA was extracted from the matching tumor and normal tissue from the same 42 patients using the Qiagen GeneRead DNA FFPE Kit protocol optimized for our FFPE slice thicknesses.

DNA Quantification and Characterization

[0416] The MNased DNA samples from the cell lines were quantified using the Quant-it Broad Range kit (Thermo Fisher Scientific, Waltham, MA) and characterized using the Bioanalyzer 1K kit (Agilent, Santa Clara, CA). To simulate cfDNA, the mononucleosomal fragments (150 bp) of each cell line were isolated via size selection and re-quantified and characterized to confirm target fragment size.

[0417] cfDNA extracted from the patient plasma was quantified using the Bioanalyzer High Sensitivity Kit (Agilent, Santa Clara, CA). The Bioanalyzer electropherograms were also used to characterize DNA fragment sizes in patient samples. The DNA extracted from the patient tissues was quantified via NanoDrop (Thermo Fisher Scientific, Waltham, MA).

Tumor Cell Line Titrations

[0418] Mononucleosomal DNA from each CNV-affected and non-affected cell line pair was purified and mixed to generate a range of known CNV titrations. These titrations simulate the stated TCFs of 1%, 2%, 3%, 5%, 7%, 10%, and 20%.

Tumor Cell Fraction

[0419] A sample with 1% TCF refers to a sample containing DNA from 1 tumor cell per 99 wild type cells. TCF as a unit was used to correctly describe titrations. TCF incorporates the increase in genome weight in strongly CNV-affected samples. Not adjusting for this genome weight gain causes unaffected regions to be present at CN<2.

Results

[0420] Table 34 provides the combined pool configuration for each region of interest. As indicated, using the above criteria, between 81 and 98% of available SNPs were selected and a pool of 4327 SNPs was selected. The primer pools included SNPs with minor allele frequencies between 0.10 and 0.50. Haploblock sizes are shown in FIG. 7. Block sizes range from 2 to 57 SNPs.

Table 34. Final pool configuration for each region of interest.

Gene	Number of assays	Expected no. heterozygous SNPs	Number of SNPs in blocks >= 10	Longest block
KRAS	493	124	254	36
EGFR	611	169	359	47
MET	539	139	345	44
FGFR1	495	126	198	47
PIK3CA	430	117	227	49
MYC	426	121	161	17
BRAF	836	231	273	57
ERBB2	497	150	236	30

Copy number truth analysis in cell lines

[0421] We observed that the Oncoscan results may not correspond to the copy numbers observed by other methods. For example, the PIK3CA region of HCC1954 and the presence of an abnormal "allele ratio from 0.5" [absolute value of (β-Allele Frequency-0.5)] that suggests a CNV beginning upstream from the copy number call made by Oncoscan.

[0422] In addition, consider for example HCC1954 in chromosome 7 (including the genes BRAF, EGFR and MET), where the BAF looked stable in Oncoscan data across the whole chromosome. According to our initial analysis, we would expect approximately 9% to 12% AAI in each of these regions for 20% TCF. However, we observed 0% AAI in BRAF and ~3.5% AAI in EGFR and MET.

[0423] Another example is the FGFR in HCC2218. Oncoscan data suggested a one copy deletion that should have resulted in 9% AAI for 20% TCF. We observed ~5% AAI for this titration.

[0424] Hence, we have decided to use the 20% TCF sample together with the Oncoscan data to determine an approximate truth for each region. More specifically, let H_1 and H_2 denote the copy numbers of the homologs, and let AAI_{20} denote the AAI found from 20% TCF sample. We only considered the regions where the average confidence on the AAI call across three replications was 99% for TCF = 20%, and we calculated the average AAI of three replications to find AAI_{20}.

[0425] We used the formula AAI = TCF*(H_1 - H_2) / [(1-TCF)*2+TCF*(H_1 + H_2)], and plug in AAI_{20}, TCF = 20%, and H_1 + H_2 found from Oncoscan analysis to estimate (H_1 - H_2). Then, we use this estimate to find the approximate expected AAI for TCF = 1%, 2%, 3%, 5%, 7%, 10%.

[0426] More specifically, the regions shown in Table 35 were found to be good candidates in each cell line (i.e., they had obvious copy number differences between the homologs). This method provided successful AAI estimations for several gene regions including HCC2218 KRAS (a non-affected region) and MET (a CNV affected region).

Table 35. Samples used in the titration experiment.

Matched Cell Line Pairs	Regions Included	Regions Excluded
HCC1954	EGFR, ERBB2, FGFR1, KRAS, MET, PIK3CA	BRAF (0% AAI in the TCF=20% sample despite clear abnormality in tumor) MYC (balanced duplication covering most of the region, with a small unbalanced duplication)
HCC2218	BRAF, ERBB2, FGFR1, MET	EGFR (CN=2 in tumor) KRAS (CN=2 in tumor) MYC (balanced duplication) PIK3CA (CN=2 in

Matched Cell Line Pairs	Regions Included	Regions Excluded
		tumor)
HCC38	BRAF, ERBB2, FGFR1, KRAS, MYC, PIK3CA	EGFR (not >99% confident about >0% AAI in the TCF=20% sample despite clear abnormality in tumor) MET (0% AAI in the TCF=20% sample despite clear abnormality in tumor)

Results of NGS analysis of geneticist tumor samples

[0427] For each tumor sample, regions with significant AAI were determined using the tumor analysis described in Kirkizlar *et al.* (Kirkizlar, Eser et al. Detection of Clonal and Subclonal Copy-Number Variants in Cell-Free DNA from Patients with Breast Cancer Using a Massively Multiplexed PCRMethodology. Translational Oncology 8.5 (2015): 407-416). Due to the mosaic nature of the tumor samples, we only aimed to determine the percentage of the SNPs affected by a CNV.

[0428] To summarize, out of 42 samples, 27 of them had at least one region with >50% of the SNPs affected and 30 of them had at least one region with >25% of the SNPs affected.

[0429] More specifically, Table 36 provides the number of samples vs. number of abnormal regions (where the abnormality is defined as >50% SNPs or >25% SNPs being affected in a region). Note that due to subclonality, the absence of CNVs in the tumor samples do not imply the absence of CNVs in the plasma.

Table 36. Summary of affected regions.

	Number of Regions with Abnormality					
	0	**1**	**2**	**3**	**4**	**>5**
>50% SNPs affected	15	9	10	2	4	2
>25% SNPs affected	12	5	11	6	4	4

Analysis using AAI method

[0430] Due to the complex and mosaic nature of CNVs present in the cell lines (confirmed with non-integer copy number calls in Oncoscan data), the titration samples with 20% TCF were used in addition to Oncoscan when determining the expected AAI in each region of each cell line. FIG. 8 provides the AAI estimate as a function of TCF values and total copy number of the tumor cells (assuming an unbalanced duplication where one homolog has one copy).

[0431] In six titrations (TCF = 1%, 2%, 3%, 5%, 7%, 10%) we found a total of 62 regions with at least 1% expected AAI across three cell lines and six titrations. Since we had three replicates at each TCF level, we made a total of 186 calls in these regions. Note that in this approach, balanced CNVs are not detected in the reference method, and so do not detract from sensitivity.

[0432] We called a region as positive if one of the following conditions satisfied: (i) AAI and confidence estimates found using all the SNPs in a region exceeded the region-level thresholds (ii) there exists a subregion with at least 50 consecutive SNPs that had AAI and confidence estimates that exceeded the subregion thresholds.

[0433] The sensitivity was 100%(51/51) for AAI ≥ 5%, 100% (60/60) for AAI ≥ 4%, 97.6% (82/84) for AAI ≥ 3%, and 91.5% (107/117) for AAI ≥ 2%. The observed specificity was 100% (336/336).

[0434] Moreover, note that TCF = 5% with a CNV of 3 copies corresponds to AAI = 2.44%.We observed that our sensitivity for AAI ≥ 2.44% was 96% (95/99).

[0435] Furthermore, specificity was determined using 24 putative normal plasma samples and six replicates of each cell line at 0% TCF, resulting in 42 * 8 = 336 regions with 0% target AAI.

[0436] The sensitivity at each expected AAI range and specificity is as in Table 37. This represents the fraction of CNVs that were successfully detected, for affected genes with expected AAI in the identified range (based on Oncoscan and 20% titration samples).

Table 37: Sensitivity and specificity at various AAI levels

Expected AAI	**TCF** **(for CN =3)**	**Called**	**Eligible**	**Sensitivity**
[1%, 2%)	[2.02%, 4.08%)	22	69	31.88%
[2%, 3%)	[4.08%, 6.19%)	25	33	75.76%
[3%, 4%)	[6.19%, 8.33%)	22	24	91.67%
[4%, 5%)	[8.33%, 10.53%)	9	9	100.00%
[5%, 8%)	[10.53%, 17.39%)	24	24	100.00%
≥ 8%	≥ 17.39%	27	27	100.00%
Expected AAI		**Called**	**Eligible**	**Specificity**
0%	0%	0	336	100.00%

[0437] Table 38 provides the sensitivity and specificity at each region in the base case scenario (note that we merged some expected AAI buckets due to the low number of regions available). Note that N/A denotes the absence of any samples within the expected AAI range under consideration. Sample size at each region is given in parentheses.

Table 38. Sensitivity, specificity, and sample size per gene at various AAI levels in the base case (Titration)

	BRAF	EGFR	ERBB2	FGFR1	KRAS	MET	MYC	PIK3CA
NumSNPs →	836	611	497	495	493	539	426	530
Target AAI	Sensitivity							
[1%, 2%)	41.67% (12)	83.33% (6)	16.67% (6)	53.33% (15)	16.67% (6)	0.00 % (9)	33.33 % (6)	0.00% (9)
2%, 4%)	100.00% (12)	N/A (0)	66.67% (9)	86.67% (15)	100.00 % (6)	50.00 % (6)	N/A (0)	77.78% (9)
≥ 4%	100.00% (3)	N/A (0)	100.00 % (36)	N/A (0)	100.00 % (3)	100.00% (9)	N/A (0)	100.00% (9)
Target AAI	Specificity							
0%	100.00% (42)	100.00 % (42)	100.00 % (42)	100.00 % (42)	100.00 % (42)	100.00% (42)	100.00% (42)	100.00% (42)

[0438] Table 39 provides the minimum expected AAI level at which all three replicates were called as positive for a given cell line and region. For regions annotated with N/A, allelic imbalance was not detected in Oncoscan and NGS. Note that no CNVs were detected at any titration level for genes that were known to be unaffected based on the reference data. This table also provides minimum AAI detected in each region (last column) and in each cell line (last row).

Table 39: Minimum expected AAI detected in all three replicates in each region and cell line in the base case (Titration)

	HCC195 4	HCC2218	HCC38	Min
BRAF	N/A	2.75%	1.80%	1.80%
EGFR	1.90%	N/A	N/A	1.90%
ERBB2	6.12%[1]	5.45%	2.96%	2.96%
FGFR1	3.01%	1.15%	1.66%	1.15%
KRAS	2.39%	N/A	2.70%	2.39%
MET	3.64%	3.14%	N/A	3.14%
MYC	N/A	N/A	2.68%	2.68%
PIK3CA	3.58%	N/A	2.72%	2.72%
Min	1.90%	1.15%	1.66%	

[1] 6.12% AAI for ERBB2 region of HCC1954 corresponded to the lowest TCF (1%) in the experiment. Hence, performance in ERBB2 is potentially closer to the one

observed in HCC38.

[0439] Next, we revised our confidence & AAI thresholds to see how the sensitivity & specificity changes. More specifically, we decreased the confidence & AAI thresholds gradually, and Table 40 below summarizes our findings.

Table 40. Sensitivity vs. specificity as a function of calling thresholds (Titration)

	Base Case	**5% lower**	**10% lower**	**15% lower**	**20% lower**	**25% lower**
Expected AAI	**Sensitivity**					
[1%,2%)	31.88%	42.03%	52.17%	57.97%	69.57%	73.91%
[2%, 3%)	75.76%	75.76%	78.79%	78.79%	81.82%	87.88%
[3%, 4%)	91.67%	91.67%	91.67%	95.83%	100.00%	100.00%
[4%, 5%)	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
[5%, 8%)	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
≥ 8%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Expected AAI	**Specificity**					
0%	100.00%	98.51%	97.32%	94.64%	93.75%	92.56%

Analysis of cell line titrations and plasmas using FODDOR method

Sample Classifier

[0440] We used FODDOR to simply classify the plasmART samples as positive/negative. The VTF is calculated from the TCF using the numbers in Table 33.

HCC38

[0441] The performance of FODDOR on this cell line is listed in Table 41. This table shows that the FODDOR sensitivity and specificity around a VTF of 5% are 100% and 97.5% respectively.

Table 41. FODDOR classifier sample level calls on cell line HCC38 at different TCF titrations.

TCF(%):	0	**0.1**	**0.2**	**1**	**2**	**3**	**5**	**10**	**20**
VTF(%):	0	0.133	0.266	1.33	2.66	3.99	6.65	13.3	26.6
Total	40	5	5	5	5	5	5	5	5

TCF(%):	0	0.1	0.2	1	2	3	5	10	20
Positive Calls	1	2	0	0	3	5	5	5	5
Negative Calls	39	3	5	5	2	0	0	0	0

HCC1954

[0442] The performance of FODDOR on this cell line is listed in Table 42. This table shows that the FODDOR sensitivity and specificity around a VTF 5% are 70% and 97.5% respectively. But notice that this cell line has multiple (in fact, all) regions with abnormal copy numbers. So, the VTF estimate is not accurate.

Table 42. FODDOR classifier sample level calls on cell line HCC1954 at different TCF titrations.

TCF(%):	0	0.1	0.2	1	2	3	5	10	20
VTF(%):	0	3.5385	7.077	35.385	70.77	106.155	176.925	353.85	707.7
Total	40	5	5	5	5	5	5	5	5
Positive Calls	1	4	3	5	5	5	5	5	5
Negative Calls	39	1	2	0	0	0	0	0	0

HCC2218

[0443] The performance of FODDOR on this cell line is listed in Table 43. This table shows that the FODDOR sensitivity even up to VTF of 10% is just 20%. This cell line is listed as having a copy number of 6.32 for the ERBB2 region. But, we noticed that this region has both deletions and duplications and so we are running into one of the limitations of the FODDOR classifier here. The specificity is 97.5%.

Table 43. FODDOR classifier sample level calls on cell line HCC2218 at different TCF titrations.

TCF(%):	0	0.1	0.2	1	2	3	5	10	20
VTF(%):	0	0.432	0.864	4.32	8.64	12.96	21.6	43.2	86.4
Total	40	5	5	5	5	5	5	5	5
Positive Calls	1	1	0	1	1	2	5	5	5
Negative Calls	39	4	5	4	4	3	0	0	0

Iterative Estimator

HCC38

[0444] Next we ran the iterative region level estimator to make calls on individual regions. At 5% TCF, the region level VTF estimates are as follows. This algorithm identifies EGFR, KRAS and MET as normal regions and the remaining regions as abnormal with VTF estimates listed in Table 44.

Table 44. FODDOR based iterative estimator's region level VTF estimates on cell line HCC38.

BRAF	EGFR	ERBB2	FGFR1	KRAS	MET	MYC	PIK3CA
0.024312	0	0.046294	0.070133	0	0	0.055685	0.016963

[0445] These results are in line with the AAI calls at 5% TCF, except for the MYC region. FODDOR identifies MYC as abnormal but AAI does not have enough confidence at 5% TCF to identify this region as abnormal. AAI successfully identifies this region as abnormal at 20% TCF as seen in Table 45. This is one of the benefits of using FODDOR in combination with AAI.

Table 45. AAI region level positive call confidences on cell line HCC38 at 5% TCF.

BRAF	EGFR	ERBB2	FGFR1	KRAS	MET	MYC	PIK3CA
99.79%	6.80%	100%	100%	57.78%	41.32%	77.73%	100%

[0446] For TCF higher than 5% FODDOR fails to make calls on any of the regions. This is because FODDOR is unable to identify a subset of at least two regions which it can use as reference regions. This seems to be the case for this cell line as can be noticed in Table 46.

Table 46. AAI region level positive call confidences on cell line HCC38 at 20% TCF.

BRAF	EGFR	ERBB2	FGFR1	KRAS	MET	MYC	PIK3CA
100%	92.12%	100%	100%	99.54%	38.21%	99.74%	100%

HCC1954

[0447] At 5% TCF, the region level VTF estimates are as follows. According to Table 31, all the regions in this cell line are abnormal. Since there are no reference regions, FODDOR is not applicable for this sample. So FODDOR results on this sample must be carefully interpreted.

[0448] First, the regions FGFR1 and KRAS have both copy deletions and so their copy number is 0. Since these regions have the least copy number, FODDOR will see these regions are normal and use them as reference regions to estimate the VTF of the other regions. In fact that is exactly what FODDOR is doing as can be seen in Table 47.

[0449] From Table 31, we see that the ERBB2 region has the highest copy number of 37. FODDOR's estimate of region level VTF's also shows that ERBB2 is the region with the highest copy number. So, even though FODDOR doesn't find a valid reference region, it is still able to pick up an extremely amplified region.

[0450] The MYC region, according to the Table 31, has a copy number of 6.59 with partial segments having balanced duplications. As a result, AAI algorithm fails to detect the abnormality in this region. Even though FODDOR doesn't have a valid reference region, we can see that FODDOR algorithm calls this regions as abnormal with a large VTF estimate. Of course, in this particular example, FODDOR was unfairly enabled to call the MYC region as abnormal. But the fact that FODDOR has a large VTF estimate for this region suggests that even if the deleted regions were actually balanced, FODDOR would have still caught the MYC abnormality. The takeaway from this example is that FODDOR is not affected by balanced CNVs and so using it in combination with AAI will enable us to catch balanced CNVs that AAI fails to catch (Table 48).

Table 47. FODDOR based iterative estimator's region level VTF estimates on cell line HCC1954

BRAF	**EGFR**	**ERBB2**	**FGFR1**	**KRAS**	**MET**	**MYC**	**PIK3CA**
0.042896	0.039741	0.50063	0	0	0.049237	0.10048	0.054556

Table 48. AAI region level positive call confidences on cell line HCC1954 at 5% TCF.

BRAF	**EGFR**	**ERBB2**	**FGFR1**	**KRAS**	**MET**	**MYC**	**PIK3CA**
23.67%	90.31%	100%	41.26%	99.98%	35.75%	53.59%	70.49%

[0451] For TCF higher than 5% FODDOR fails to make calls on any of the regions. This is because FODDOR is unable to identify a subset of at least two regions which it can use as reference regions. This, in fact seems to be the case for this cell line as can be noticed in the Table 49.

Table 49. AAI region level positive call confidences on cell line HCC1954 at 20% TCF.

BRAF	**EGFR**	**ERBB2**	**FGFR1**	**KRAS**	**MET**	**MYC**	**PIK3CA**
26.56%	100%	100%	100%	100%	100%	54.79%	100%

HCC2218

[0452] At 5% TCF, the region level VTF estimates are in Table 50.

Table 50. FODDOR based iterative estimator's region level VTF estimates on cell line HCC2218

BRAF	EGFR	ERBB2	FGFR1	KRAS	MET	MYC	PIK3CA
0.024168	0	0	0	0.0058011	0.038968	0.061569	0.013419

[0453] According to the Table 31 only the EGFR and KRAS regions are normal. The FGFR1 has a deletion resulting in a copy number of 1. But since FODDOR assumes the regions with the least copy number as normal, it incorrectly sees FGFR1 as normal and as a consequence sees KRAS and PIK3CA as abnormal. In this cell line, the ERBB2 region has both deletion and duplication which effectively canceled each other and so FODDOR identified this region as normal.

[0454] Next, we removed the FGFR1 region from the analysis and re-ran the FODDOR based iterative estimator on the rest of the regions. Since there are no deletions in any of the other regions, we expected FODDOR to perform correctly. The new VTF estimates are in Table 51.

Table 51. FODDOR based iterative estimator's region level VTF estimates on cell line HCC2218 after eliminating the FGFR1 region.

BRAF	EGFR	ERBB2	FGFR1	KRAS	MET	MYC	PIK3CA
0	0	0	N/A	0	0.0197	0.0459	0.0045

[0455] Now, the results look reasonable. FODDOR correctly identified MYC and MET as abnormal regions. In 1 out of 5 replicates, it identified PIK3CA as abnormal. Also, it correctly called EGFR and KRAS as normal. FODDOR failed to detect the BRAF abnormality. The ERBB2 abnormality was again not detected as expected due to the reason previously explained. You can compare FODDOR results against the AAI results listed in Table 52.

Table 52. AAI region level positive call confidences on cell line HCC2218 at 5% TCF.

BRAF	EGFR	ERBB2	FGFR1	KRAS	MET	MYC	PIK3CA
100%	70.95%	90.49%	95.64%	8.12%	100%	47.89%	43.81%

[0456] Here again, the MYC region has balanced duplications. As a result, the AAI algorithm failed to detect the abnormality in this region. But FODDOR algorithm was able to successfully detect this abnormality. Notice that AAI fails to detect the MYC abnormality even at 20% TCF as shown in Table 53.

Table 53. AAI region level positive call confidences on cell line HCC2218 at 20% TCF.

BRAF	EGFR	ERBB2	FGFR1	KRAS	MET	MYC	PIK3CA
100%	81.73%	100%	100%	74.46%	100%	58.7%	22.96%

Stand alone FODDOR performance

[0457] Here, we analyzed the stand-alone performance of the FODDOR based iterative estimator. We used the AAI calls at 20% TCF as the truth and compared the FODDOR results to this truth. Since the AAI has some limitations, using the AAI results as the truth to estimate FODDOR performance, especially the sensitivity, gave us the lower limit of the true performance.

HCC38

[0458] Here, we know from AAI that ERBB2 is abnormal, but since we do not have the copy number estimate of this region we assumed ERBB2 was normal for this analysis. The individual region copy numbers that we used for computing the VTF from the TCF are obtained from Table 31. The results are shown in Table 54.

Table 54. FODDOR performance on cellline HCC38 using AAI calls at 20% TCF as truth.

VTF	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**
Positives	45	25	15	5	5	5	0
True Positives	22	17	13	5	5	5	0
Sensitivity	48.889%	68%	86.667%	100%	100%	100%	NaN

HCC1954

[0459] Here we assumed that FGFR1 and KRAS are normal and analyzed the calls on the rest of the regions. The results are shown in Table 55.

Table 55. FODDOR performance on cellline HCC1954 using AAI calls at 20% TCF as truth.

VTF >=	0.01	0.02	0.03	0.04	0.05	0.06	0.07
Positives	125	105	85	65	55	40	40
True Positives	72	71	66	57	52	37	37
Sensitivity	57.6%	67.619%	77.647%	87.692%	94.545%	92.5%	92.5%

HCC2218

[0460] Here we assumed that ERBB2 is normal because there is both deletion and duplication on this region which effectively makes the abnormality invisible to FODDOR. The results are shown in Table 56.

Table 56. FODDOR performance on cellline HCC2218 using AAI calls at 20% TCF as truth.

VTF	0.01	0.02	0.03	0.04	0.05	0.06	0.07
Positives	55	45	30	20	15	5	5
True Positives	17	16	16	11	11	5	5
Sensitivity	30.909%	35.556%	53.333%	55%	73.333%	100%	100%

[0461] Combining the results from the three cell line samples with a VTF estimate of at least 5%, we estimate the sensitivity of FODDOR region level estimator to be 92.35%.

[0462] In addition to the above titration samples, we also ran FODDOR on 120 wild type samples. One these FODDOR made positive calls on 2 samples (2 regions on one and 5 regions on the other). 2 other samples had extremely bad K-S statistics suggesting that something unusual was happening with theses samples. So, we removed these 2 samples from our analysis. Using these numbers the estimated specificity of 99.26%.

Summary

[0463] The FODDOR based iterative estimator by itself has lower performance than AAI. But when combined with AAI, it can improve the overall performance by detecting abnormalities that AAI fails to detect, due to fundamental limitations of AAI.

Sample QC

[0464] As a part of sample QC, we determined the following:

(1) Match between the tumor and plasma samples (to make sure that the plasma sample is coming from the same person whose tumor we analyzed)

2) Contamination checks to determine the presence of ambient or genotyped contamination. We used a tentative 0.2% threshold to determine if a sample is contaminated or not.

[0465] Based on this analysis, we observed that two cancer plasma samples, namely 9770Vd(303) and 9545 VH with sequencing ids 2330093 and 2330135, did not match the genotypes of their corresponding tumor sample and were eliminated from further analysis (note that to ensure that the abnormality in the tumor is not causing the mismatch, we only looked at the heterozygous SNPs when making this determination).

[0466] Furthermore, cancer plasmas 2872/12 and 5679/12 with sequencing ids 2330110 and 2330114, had high level of genotyped contamination (>2%) and were also eliminated from further analysis.

[0467] Two negative samples (Neg-9 and Neg-37 with sequencing ids 2330145 and 2370513), were mixtures of multiple plasmas. One other negative sample (Neg-22 with sequencing id 2370501) had higher than usual contamination (0.3% ambient and 0.9% genotyped contamination). These three samples were also eliminated from further analysis. Hence, a total of 4 cancer and 3 negative plasmas failed the QC, resulting in remaining 38 cancer and 83 negatives for the further analysis.

Analysis using AAI method

[0468] Sample level calls using the base case thresholds using in the titration analysis were as in Table 57. Note that the sensitivity and specificity is not exactly well defined in this context, although we still use these terms loosely. More specifically, we detected 34.21% of all cancer samples as positives. Moreover, one normal sample was called as positive with 100% confidence and an AAI of 4.22% (Neg-91 with sequencing id 2370539). This sample actually seems to have an abnormality that is visible in the het rate plot hence we believe it is very likely to be a correct call for analytical purposes.

Table 57. Sample level sensitivity vs. specificity in the base case (Geneticist)

Stage	**Called**	**Eligible**	**Sensitivity**
IA	3	8	37.50%
IB	7	19	36.84%
IIB	3	6	50.00%
IIIA	0	4	0.00%
IV	0	1	0.00%
All Cancer	13	38	34.21%
	Called	**Eligible**	**Specificity**
Normal	1	83	98.80%

[0469] Next, we provide the sensitivity in the samples with positive tumors. More specifically, we define a tumor as positive if there is at least one sample with at least 50% SNPs covered with a CNV based on the previous analysis. Table 58 below provides the sensitivity for plasmas with positive and negative tumors.

Table 58. Sample level sensitivity as a function of presence of CNVs in the tumor (Geneticist)

	TumorPos	**TumorNeg**
PlasmaPos	50.00%	7.14%
PlasmaNeg	50.00%	92.86%
NumSamples	24	14

[0470] Note that the sensitivity seems significantly higher in the plasmas with corresponding positive tumors (50.00%) compared to plasmas with negative tumors (7.14%).

[0471] Finally, in Table 59 we study the sensitivity vs. specificity tradeoff as a function of calling thresholds.

Table 59. Sample level sensitivity vs. specificity as a function of calling thresholds (Geneticist)

	Base Case	**5% lower**	**10% lower**	**15% lower**	**20% lower**	**25% lower**
Stage	**Sensitivity**					
IA	37.50%	37.50%	50.00%	50.00%	50.00%	50.00%
IB	36.84%	36.84%	36.84%	36.84%	42.11%	42.11%
IIB	50.00%	50.00%	50.00%	66.67%	66.67%	66.67%
IIIA	0.00%	0.00%	0.00%	0.00%	25.00%	25.00%
IV	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
All Cancer	34.21%	34.21%	36.84%	39.47%	44.74%	44.74%
Normal	**Specificity**					
	98.80%	97.59%	97.59%	96.39%	93.98%	91.57%

Analysis using FODDOR method

[0472] First we ran the FODDOR classifier to simply classify a sample as positive/negative. The performance of the classification had a sensitivity estimate of 42.105% (38 positives, 16 true positive calls) and specificity estimate of 95.213% (188 negatives, 9 false positive calls). Note that a sample marked as positive may not necessarily have a CNV in the plasma. Also, we do not know the TCF in the plasma sample.

[0473] This specificity estimate is very close to our specificity estimate using the zero-titration samples before. Here FODDOR identified the following 7 additional samples that were not identified as positive by AAI. These samples demonstrate the benefit of using FODDOR in combination with AAI. It is important to note that FODDOR and AAI identified different cancer samples as positives. So together they identified 21 samples as positives which gives us a combined sensitivity estimate of 60.53%

[0474] Next we ran the region level estimator. As truth we used the individual region copy numbers determined previously in the CNV-truth Geneticist samples. The sensitivity estimate is 16.176% (68 positives, 11 true positive calls) and the specificity estimate is 97.065% (1772 negatives, 52 false negative calls). Note that the abnormal region in the tumor does not mean that the region is abnormal in the plasma. So, the sensitivity estimate above is only a lower bound of the true sensitivity. Also, note that we do not know the TCF here.

[0475] Using the AAI calls on individual regions, the sensitivity estimate is 31.579% (38 positives, 12 true positives) and the specificity estimate is 97.17% (1802 negatives, 51 false positives).

[0476] The sensitivity estimate above is only a lower bound of the true sensitivity.

Comparison to perfomance objectives

[0477] The performance objectives for fCNV technology were as follows, measured as analytical performance per gene region tested: (1) Sensitivity ≥ 95% for TCF ≥ 5% assuming copy number change ≥ 1 and (2)

Specificity ≥ 99%.

[0478] The sensitivity objective can be evaluated with respect to the AAI method or the combination. The AAI method detected 96% (95/99) of regions with AAI corresponding to a one copy change at TCF ≥ 5%, for regions where an allelic imbalance was present in the reference data. There were two genes with balanced CNVs, which if included in the sensitivity calculation, reduce the AAI sensitivity to 90% (95/105). These were both detectable by FODDOR, leading to a combined method sensitivity of 96% (101/105).

[0479] Specificity can be observed from real plasma that is assumed to be unaffected by a CNV because it was collected from subjects presumed to be healthy, but there is still some risk of a CNV being present. Specificity could also be estimated from samples prepared from pure wild-type cell line. Thus the observed specificity might be an underestimate of the true specificity.

[0480] The specificity of the AAI method in presumed-negative plasma samples was 98.8% (82/83) by sample or 99.8% (663/664) by region, but the sample called positive was confirmed to have an allelic imbalance visible by inspection. Therefore the estimated analytical specificity could be considered 100%. The specificity in the cell line titrations was also 100%. The specificity demonstrated by the FODDOR method was 95% in real plasma and 98% in pure wild-type cell lines.

Performance on affected patient samples

[0481] The following conclusions can be drawn: (1) samples that do not show an allelic imbalance in the tumor tissue are far less likely to show one in plasma; (2) a significant number of samples are identified as positive using one algorithm but not the other, bidirectionally; and

(3) some samples are not identified as positive by either algorithm, even conditioned on the presence of allelic imbalance in the tumor tissue.

The fact that the two algorithms identify different sets of positive samples is expected due to their differing methods and could indicate that balanced CNVs are more common than expected.

Discussion

[0482] The results of this study are equivalent to detecting a CNV with copy number of 6 and a TCF of 2%, with 100% sensitivity and 100% specificity in liquid biopsies.

[0483] To put our results into context with published results, Lanman et al. 2015 (PLoS ONE 10(10): e0140712. doi:10.1371/journal.pone.0140712) (Guardant Health) shows plasma fCNV limit of detection of 5% TCF with a copy number of 6; this means that they are able to detect an AAI of ~9.1% (More specifically, the limit of detection mentioned is an additional 0.2 copies in EGFR and MET. This corresponds to an AAI of 0.2/2.2 = 9.09%. For ERBB2 the limit is higher at 0.5 copies, or an AAI of 0.5/2.5 = 20%.) compared to the observed 100% sensitivity at AAI of 4% demonstrated in this Example.

[0484] In FIG. 8, example AAI values that were detected with 100% sensitivity using the fCNV method herein are marked with a dot pattern, and the limit of detection claimed by Lanman et al. is marked with a line pattern.

REFERENCES CITED IN THE DESCRIPTION

Cited references

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- WO2015164432A [0007] [0139]
- WO2007062164A [0139]

- WO2012108920A [0139]
- US20120270212A [0180]
- US20120185176A [0207]
- US20140065621A [0207]
- US20120264121A [0208]
- US7754428B [0208]
- US7901884B [0208]
- US8166382B [0208]
- US20120190020A [0239]
- US20120190021A [0239]
- US20120190557A [0239]
- US20120191358A [0239]

Non-patent literature cited in the description

- **KALNINA et al.**World J Gastroenterol., 2015, vol. 21, 4111636-11653 [0064]
- **JIANG et al.**Proc Natl Acad Sci USA, vol. 112, E1317-E1325 [0064]
- **HAMAKAWA et al.**Br J Cancer., 2015, vol. 112, 352-356 [0065]
- **UNTERGRASSER ACUTCUTACHE IKORESSAAR TYE JFAIRCLOTH BCREMM MROZEN SG**Primer3 - new capabilities and interfaces.Nucleic Acids Research, 2012, vol. 40, 15e115- [0082]
- **KORESSAAR TREMM M**Enhancements and modifications of primer design program Primer3.Bioinformatics, 2007, vol. 23, 101289-91 [0082]
- **SANTALUCIA JR**A unified view of polymer, dumbbell and oligonucleotide DNA nearest-neighbor thermodynamicsProc Natl Acad Sci, 1998, vol. 95, 1460-65 [0082] [0093] [0300]
- **JAMES T. ROBINSONHELGA THORVALDSDÓTTIRWENDY WINCKLERMITCHELL GUTTMANERIC S. LANDERGAD GETZJILL P. MESIROV**Integrative Genomics ViewerNature Biotechnology, 2011, vol. 29, 24-26 [0084]
- **KENT WJSUGNET CWFUREY TSROSKIN KMPRINGLE THZAHLER AMHAUSSLER D**The human genome browser at UCSCGenome Res., 2002, vol. 12, 6996-1006 [0084]
- Integrated genomic analysis of ovarian carcinomaNature, 2011, vol. 474, 609-616 [0114]
- **SU et al.**J Mol Diagn, 2011, vol. 13, 74-84 [0128]
- **ABAAN et al.**The Exomes of the NCI-60 Panel: A Genomic Resource for Cancer Biology and Systems PharmacologyCancer Research, 2013, [0128]
- **CALIN et al.**A microRNA signature associated with prognosis and progression in chronic lymphocytic leukemia.N Engl J Med, 2005, vol. 353, 1793-801 [0128]
- **RYAN et al.**A prospective study of circulating mutant KRAS2 in the serum of patients with colorectal neoplasia: strong prognostic indicator in postoperative follow upGut,

2003, vol. 52, 101-108 [0132]

- **LECOMTE T et al.**Detection of free-circulating tumor-associated DNA in plasma of colorectal cancer patients and its association with prognosisInt J Cancer, 2002, vol. 100, 542-548 [0132]
- **CHEN et al.**Nat. Rev. Cancer., 2014, vol. 14, 8535-551 [0133]
- **BROWNINGBROWNING**Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype ClusteringAm J Hum Genet, 2007, vol. 81, 51084-1097 [0169]
- **STEPHENSSCHEET**Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing-Data ImputationAm. J. Hum. Genet., 2005, vol. 76, 449-462 [0170]
- **SCHEETSTEPHENS**A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.Am J Hum Genet, 2006, vol. 78, 629-644 [0171]
- **HOWIEDONNELLYMARCHINI**A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.PLoS Genetics, 2009, vol. 5, 6e1000529- [0172]
- **DELANEAUCOULONGESZAGURY**Shape-IT: new rapid and accurate algorithm for haplotype inferenceBMC Bioinformatics, 2008, vol. 9, 540- [0173]
- **ERONENGEERTSTOIVONEN**HaploRec: Efficient and accurate large-scale reconstruction of haplotypesBMC Bioinformatics, 2006, vol. 7, 542- [0174]
- **QINNIULIU**Partition-Ligation-Expectation-Maximization Algorithm for Haplotype Inference with Single-Nucleotide PolymorphismsAm J Hum Genet., 2002, vol. 71, 51242-1247 [0175]
- **KIMMELSHAMIR**GERBIL: Genotype Resolution and Block Identification Using LikelihoodProceedings of the National Academy of Sciences of the United States of America (PNAS), 2005, vol. 102, 158-162 [0176]
- **CLAYTON, D**SNPHAP: A Program for Estimating Frequencies of Large Haplotypes of SNPs, 2002, [0177]
- **BRINZAZELIKOVSKY**2SNP: scalable phasing based on 2-SNP haplotypesBioinformatics, 2006, vol. 22, 3371-3 [0178]
- **SHERRY STWARD MHKHOLODOV M et al.**dbSNP: the NCBI database of genetic variationNucleic Acids Res., 2001, vol. 29, 1308-11 [0180]
- **SPARKS et al.**Noninvasive prenatal detection and selective analysis of cell-free DNA obtained from maternal blood: evaluation for trisomy 21 and trisomy 18Am J Obstet Gynecol, 2012, vol. 206, 319.e1-9 [0208]
- Detection Theory**KAY S.M.**Fundamentals of Statistical Signal ProcessingPrentice-Hall, Inc.19980000vol. 2, [0252]
- **MCVEAN et al.**An integrated map of genetic variation from 1,092 human genomesNature, 2012, vol. 491, 56-65 [0299]
- **KIRKIZLAR et al.**Translational Oncology, vol. 8, 407-416 [0316] [0408]
- TCGA 2012 - 178 SQCC (Lung Squamous Cell Carcinoma) samplesNature, 2012, vol. 489, 519-25 [0354]
- TCGA 2014 - 230 ADC (Lung Adenocarcinoma) samplesNature, 2014, vol. 511, 543-

50 [0354]

- **GEORGE et al.**2015 - 110 SCLC (Small Cell Lung Cancer) samplesNature, 2015, vol. 524, 47-53 [0354]
- **KIRKIZLAR, ESER et al.**Detection of Clonal and Subclonal Copy-Number Variants in Cell-Free DNA from Patients with Breast Cancer Using a Massively Multiplexed PCRMethodologyTranslational Oncology, 2015, vol. 8, 5407-416 [0427]

FREMGANGSMÅDER TIL BESTEMMELSE AF PLOIDI

Patentkrav

1. En computerimplementeret fremgangsmåde til bestemmelse af ploidi for et kromosomsegment i en prøve fra et individ, hvilken fremgangsmåde omfatter:

a. modtagelse af allelfrekvensdata for hver SNP af et sæt af SNP'er omfattende 200 SNP'er på en flerhed af undersegmenter inde i kromosomsegmentet, hvor inde i hvert undersegment 95 % af parvise SNP-sammenligninger mellem hvilke som helst to SNP'er inde i dette kromosom/område har |D'| på >95 %, hvor allelfrekvensdataene omfatter mængden af hver allel til stede i prøven ved hver SNP;

b. generering af fasevise allelinformationer for sættet af SNP'er ved estimering af fasen af de genotypiske måledata, idet der tages højde for en øget statistisk korrelation af SNP'er inde i det samme undersegment;

c. generering af individuelle sandsynligheder for allelfrekvenser for sættet af SNP'er for forskellige ploiditilstande ved hjælp af allelfrekvensdataene;

d. generering af fælles sandsynligheder for sættet af forbundne SNP'er ved hjælp af de individuelle sandsynligheder og de fasevise allelinformationer; og

e. udvælgelse, baseret på de fælles sandsynligheder, af en model, der passer bedst, og som indikerer kromosomal ploidi, hvorved kromosomsegmentets ploidi bestemmes.

2. Fremgangsmåden ifølge krav 1, hvor

trin c endvidere omfatter oprettelse, på en computer, af et sæt af ploiditilstandshypoteser, hvor hver ploiditilstandshypotese er én mulig ploiditilstand for kromosomsegmentet;

trin d endvidere omfatter opbygning af et sæt af fælles fordelingsmodeller for forventede genotypiske målinger ved sættet af SNP'er for hver hypotese, idet identificerede kromosomovergangsplaceringer tages i betragtning; og

trin e endvidere omfatter udvælgelse af ploiditilstanden med den største sandsynlighed.

3. Fremgangsmåden ifølge krav 1 eller 2, hvor allelfrekvensdataene genereres ved måling af signalstyrker for forskellige alleler ved hjælp af et SNP-microarray; eller

modtagelse af allelfrekvensdata omfatter modtagelse af nukleinsyresekvenseringsdata for mindst 200 forskellige amplikoner, der spænder over hver SNP fra sættet af SNP'er, og generering af allelfrekvensdataene fra sekvenseringsdataene;

fortrinsvis hvor sekvenseringsdataene udledes fra sekvensering med høj gennemstrømning.

4. Fremgangsmåden ifølge et hvilket som helst af kravene 1 til 3, og som endvidere omfatter amplifikation af sættet af SNP'er ved hjælp af en amplifikationsfremgangsmåde, der omfatter:

i. dannelse af en reaktionsblanding omfattende cirkulerende frie nukleinsyrer afledt fra prøven, en polymerase og en primerpulje, der mindst omfatter 200 primere eller primerpar, som hver(t) specifikt binder til en primerbindingssekvens placeret i en effektiv afstand fra én af SNP'erne; og

ii. udsættelse af reaktionsblandingen for amplifikationsbetingelser,

hvorved der genereres af en flerhed af amplikoner; og

udsættelse af hver af amplikonerne for en nukleinsyresekvenseringsreaktion for at generere nukleinsyresekvenseringsdataene til amplikonerne;

fortrinsvis hvor amplifikationsfremgangsmåden er en PCR-reaktion og annealing-temperaturen er mellem 1 og 15 °C højere end smeltetemperaturen for mindst 50 % af primerne fra primersættet.

5. Fremgangsmåden ifølge et hvilket som helst af kravene 1 til 4, hvor der anvendes en beta-binomialfordeling til at bestemme individuelle sandsynligheder for allelfrekvenser for SNP'erne for forskellige ploiditilstande.

6. Fremgangsmåden ifølge et hvilket som helst af kravene 1 til 5, hvor der beregnes en gennemsnitlig allel-ubalance, og hvor bestemmelsen af kopiantal indikerer en kopinummervariation, hvis den gennemsnitlige allel-ubalance er på eller større end 0,45 %.

7. Fremgangsmåden ifølge et hvilket som helst af kravene 1 til 5, hvor trin e omfatter bestemmelse af ploiditilstanden med den største sandsynlighed baseret på bayesiansk estimering som en indikation af antallet af kromosomsegmentets kopier.

8. Fremgangsmåden ifølge et hvilket som helst af kravene 1 til 7, hvor sættet af SNP'er omfatter 1.000 SNP'er.

9. Fremgangsmåden ifølge et hvilket som helst af kravene 1 til 8, og som endvidere omfatter korrigering af allelfrekvensdataene for bias, kontaminering og/eller sekvenseringsfejl og opnåelse af tidligere sandsynligheder for hver hypotese fra populationsdata og beregning af konfidens ved hjælp af Bayes regel, hvor konfidens beregnes for bestemmelsen af kopiantallet.

10. Fremgangsmåden ifølge et hvilket som helst af kravene 1 til 9, og som endvidere omfatter bestemmelse af ploiditilstanden ved hjælp af en kvantitativ, ikke-allel-fremgangsmåde, hvor identifikation af kromosomsegmentet som aneuploid ved hjælp af den kvantitative, ikke-allel-fremgangsmåde eller ved hjælp af fremgangsmåden ifølge de foregående krav indikerer en kopiantalsvariation for kromosomsegmentet.

11. Fremgangsmåden ifølge krav 10, hvor den kvantitative, ikke-allel-fremgangsmåde omfatter:

a. måling af mængden af genetisk materiale af en flerhed af kromosomsegmenter;

b. sammenligning af de målte mængder af genetisk materiale for hvert af kromosomsegmenterne mod hinanden; og

c. detektering af kopiantalsvariation (CNV) eller aneuploidi ved identificering af tilstedeværelse eller fravær af en deletion eller duplikation af mindst ét af kromosomsegmenterne baseret på sammenligningen.

12. Fremgangsmåden ifølge et hvilket som helst foregående krav, hvor prøven er fra et målindivid, hvor målindividet er et født individ eller et ufødt foster.

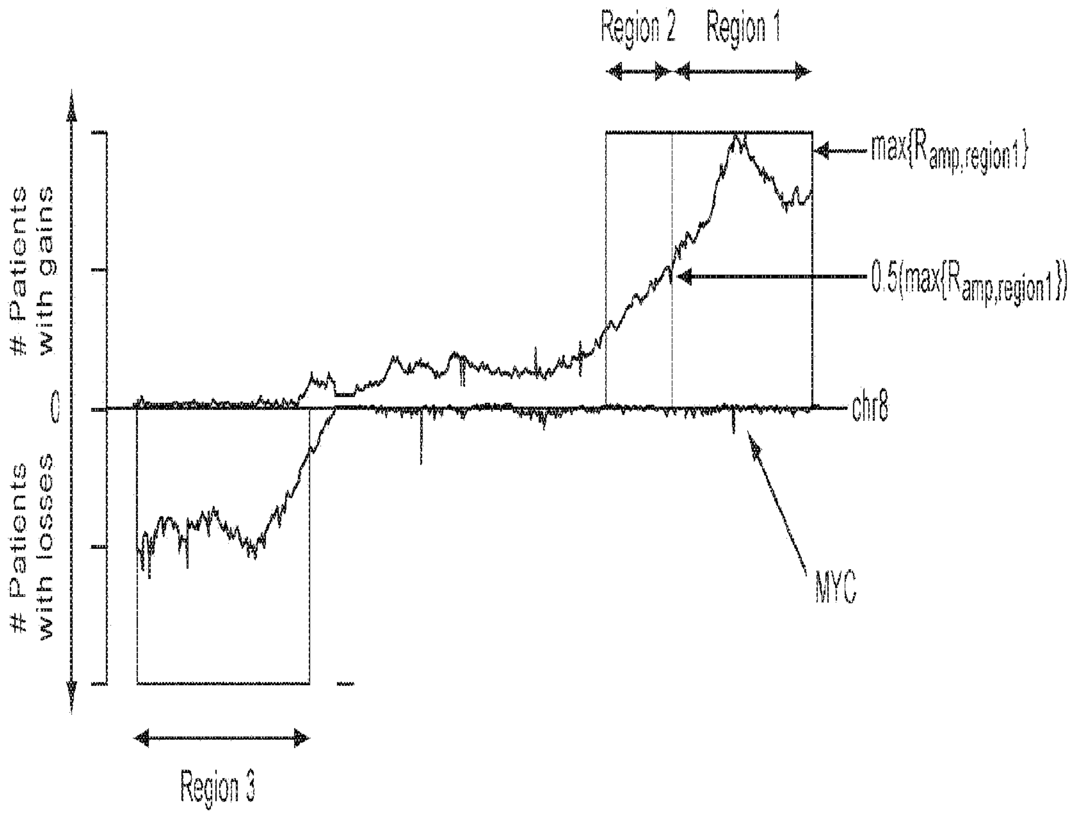
13. Fremgangsmåden ifølge et hvilket som helst af kravene 1-11, hvor prøven er fra et væv eller organ mistænkt for at have en deletion eller duplikation, såsom celler eller en masse mistænkt for at være angrebet af cancer.

14. Fremgangsmåden ifølge et hvilket som helst af kravene 1-11, hvor fremgangsmåden til non-invasiv, prænatal test og prøven omfatter celler, cfDNA eller cfRNA fra en blodprøve, eller en fraktion deraf, fra en gravid kvinde.

DRAWINGS

Drawing

FIG. 1



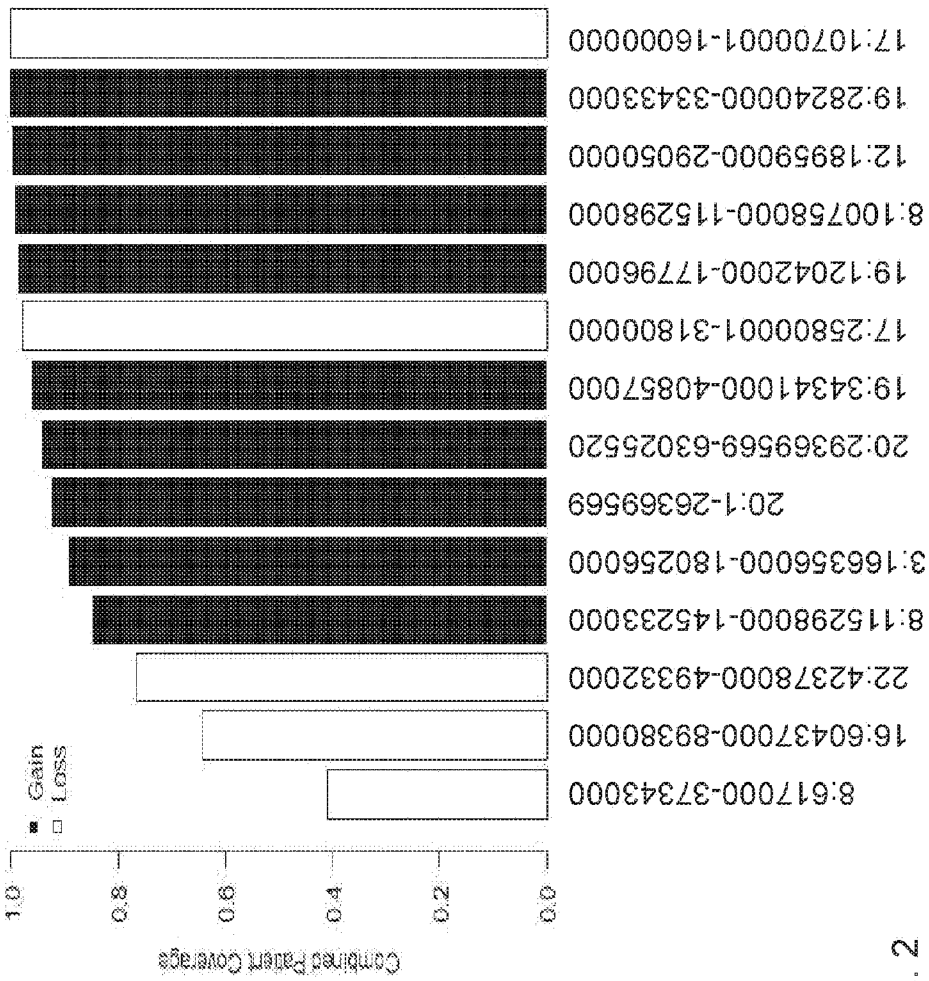

Combined Patient Coverage
1.0
0.8
0.6
0.4
0.2
0.0
■ Gain
□ Loss
8:617000-37343000
16:60437000-89380000
22:42378000-49332000
8:115298000-145233000
3:166356000-180256000
20:1-26369569
20:29369569-63025520
19:34341000-40857000
17:25800001-31800000
19:12042000-17796000
8:100758000-115298000
12:18959000-29050000
19:28240000-33433000
17:10700001-16000000

FIG. 2

Region	ID	1	2	3	4	5	6	7
12:18959000-29050000	1	**1.0000**	0.0001	0.0058	-0.0295	0.0374	0.0351	-0.0037
16:60437000-89380000	2	0.0001	**1.0000**	0.0083	0.0227	-0.0559	-0.0241	-0.0539
19:12042000-17796000	3	0.0058	0.0083	**1.0000**	**0.1696**	**0.1402**	**0.0932**	0.0407
19:28240000-33433000	4	-0.0295	0.0227	**0.1696**	**1.0000**	**0.3144**	**0.0765**	0.0366
19:34341000-40857000	5	0.0374	-0.0559	**0.1402**	**0.3144**	**1.0000**	0.0256	0.0676
22:42378000-49332000	6	0.0351	-0.0241	**0.0932**	**0.0765**	0.0256	**1.0000**	-0.0343
3:166356000-180256000	7	-0.0037	-0.0539	0.0407	0.0366	**0.0676**	-0.0343	1.0000
8:617000-37343000	8	0.0278	-0.0766	0.0134	**0.0821**	**0.0797**	0.0461	-0.0074
8:115298000-145233000	9	-0.0131	0.0066	-0.0323	-0.1370	-0.1582	-0.0688	-0.0038
8:100758000-115298000	10	-0.0339	-0.0167	-0.0145	-0.0331	-0.0847	-0.0499	0.0019
20:1-26369569	11	**0.0544**	-0.0275	0.0020	-0.0079	-0.0295	-0.0446	-0.0892
20:29369569-63025520	12	**0.0790**	0.0062	**0.0732**	**0.0747**	-0.0435	0.0259	**0.0530**
17:25800001-31800000	13	**0.0592**	0.0199	**0.0534**	**0.0518**	-0.0027	0.0115	-0.0183
17:10700001-16000000	14	**0.0829**	0.0280	-0.0124	-0.0126	-0.0633	-0.0590	0.0049

FIG. 3A

Region	I	9	10	11	12	13	14
12:18959000-29050000	1	-0.0131	-0.0339	**0.0544**	**0.0790**	**0.0592**	**0.0829**
16:60437000-89380000	2	0.0066	-0.0167	-0.0275	0.0062	0.0199	0.0280
19:12042000-17796000	3	-0.0323	-0.0145	0.0020	**0.0732**	**0.0534**	-0.0124
19:28240000-33433000	4	-0.1370	-0.0331	-0.0079	**0.0747**	**0.0518**	-0.0126
19:34341000-40857000	5	-0.1582	-0.0847	-0.0295	-0.0435	-0.0027	-0.0633
22:42378000-49332000	6	-0.0688	-0.0499	-0.0446	0.0259	0.0115	-0.0590
3:166356000-180256000	7	-0.0038	0.0019	-0.0892	**0.0530**	-0.0183	0.0049
8:617000-37343000	8	0.0493	-0.0221	-0.0264	0.0304	0.0197	0.0091
8:115298000-145233000	9	**1.0000**	**0.3285**	-0.0220	0.0271	-0.0701	-0.0733
8:100758000-115298000	10	**0.3285**	**1.0000**	-0.0488	0.0481	-0.0088	-0.0487
20:1-26369569	11	-0.0220	-0.0488	**1.0000**	-0.0156	-0.0487	0.0117
20:29369569-63025520	12	0.0271	0.0481	-0.0156	**1.0000**	0.0251	**0.0513**
17:25800001-31800000	13	-0.0701	-0.0088	-0.0487	0.0251	**1.0000**	0.0396
17:10700001-16000000	14	-0.0733	-0.0487	0.0117	**0.0513**	0.0396	**1.0000**

FIG. 3B

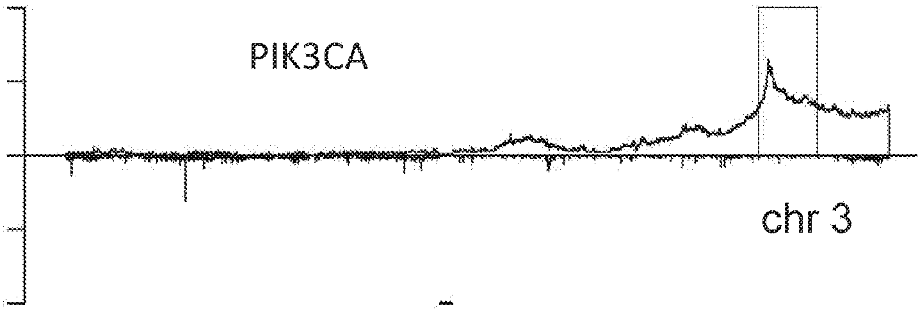


FIG. 4A

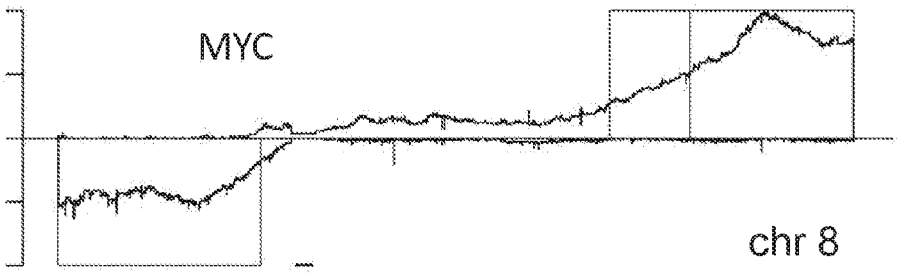


FIG. 4B

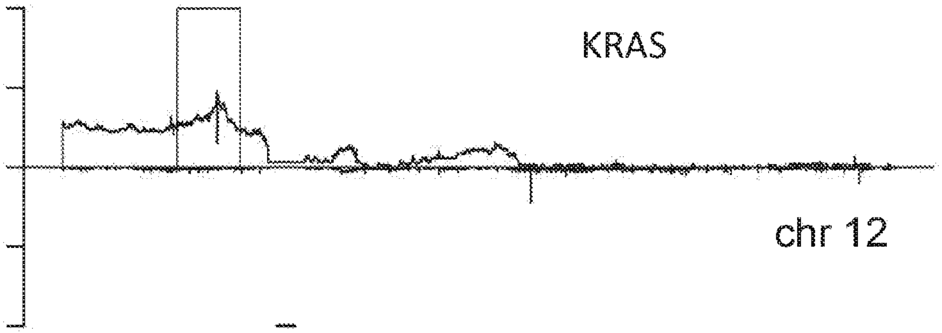


FIG. 4C

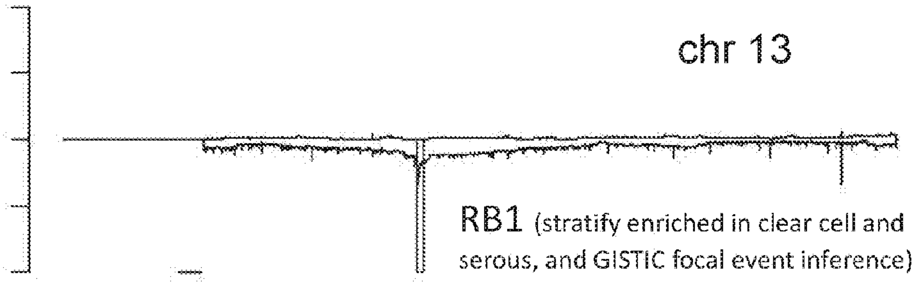


FIG. 4D

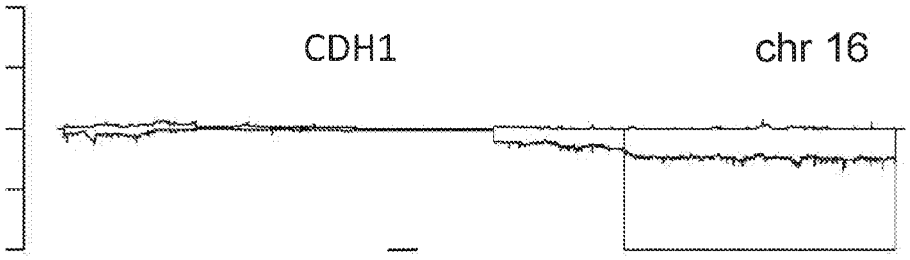


FIG. 4E

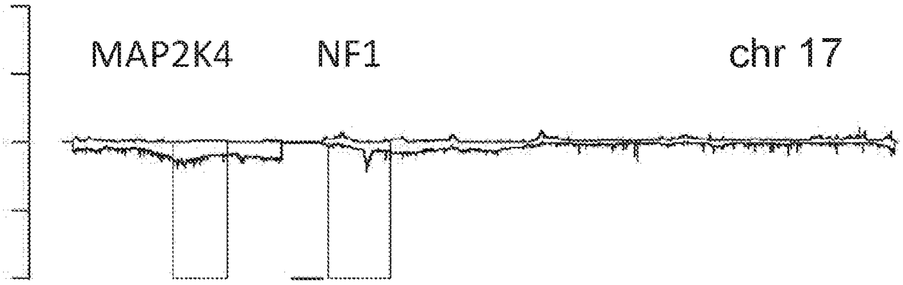


FIG. 4F

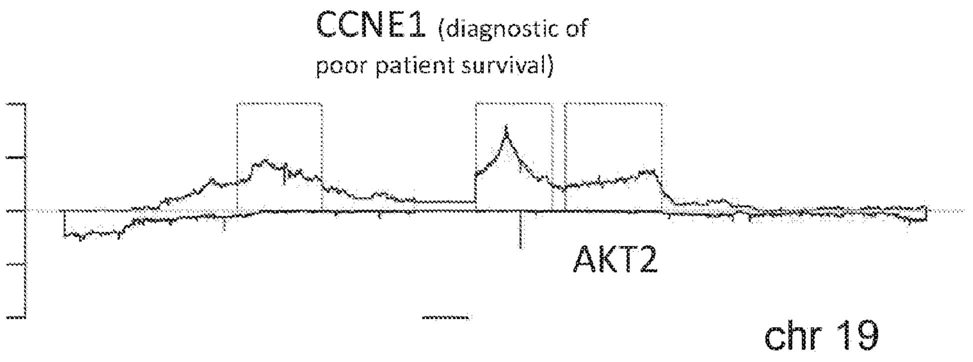

CCNE1 (diagnostic of poor patient survival)
AKT2
chr 19

FIG. 4G

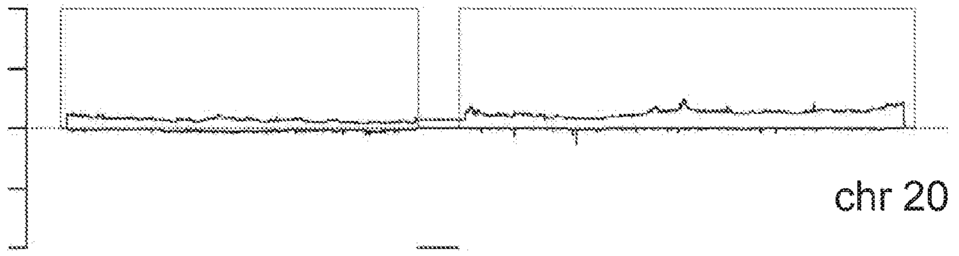

chr 20

FIG. 4H

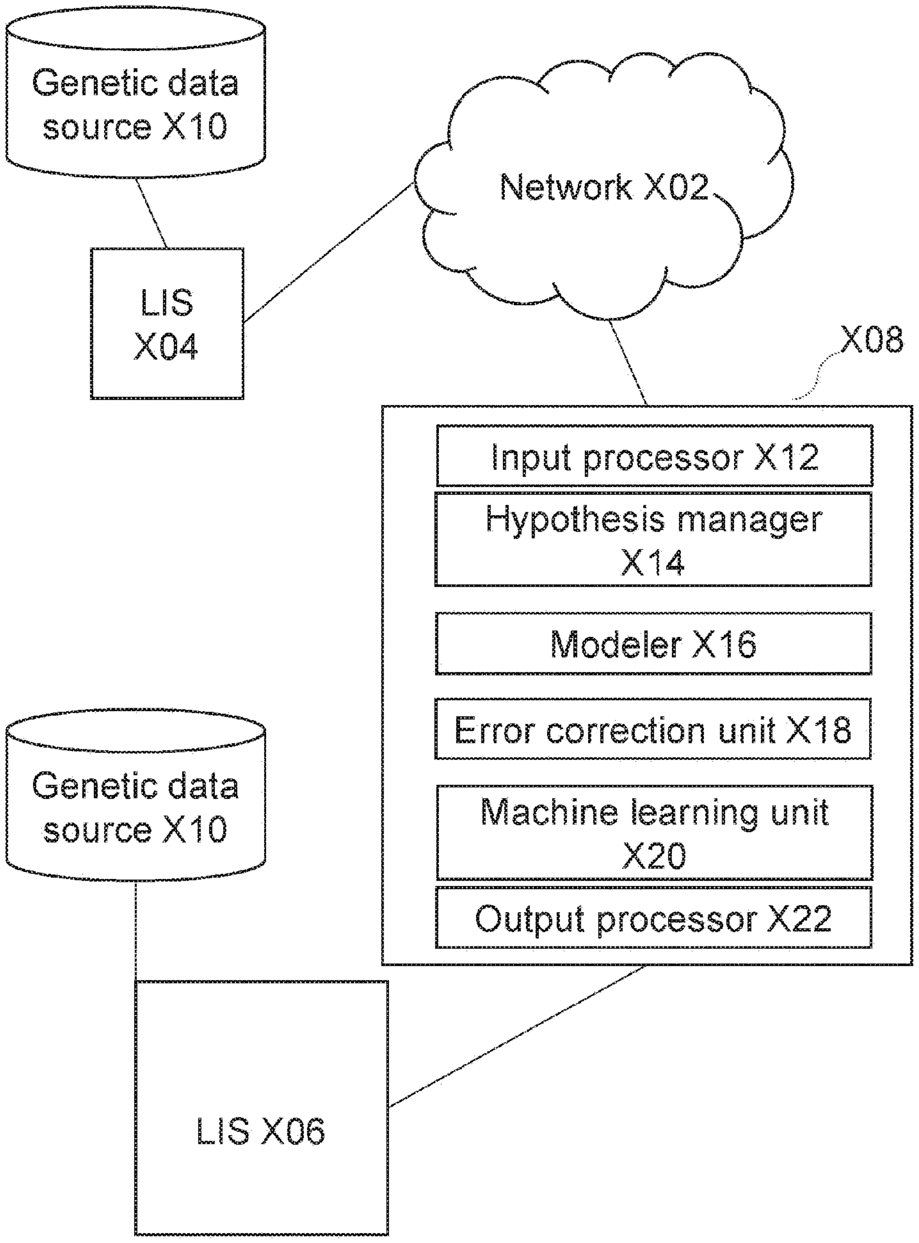

Genetic data source X10
Network X02
LIS X04
X08
Input processor X12
Hypothesis manager X14
Modeler X16
Error correction unit X18
Genetic data source X10
Machine learning unit X20
Output processor X22
LIS X06

FIG. 5

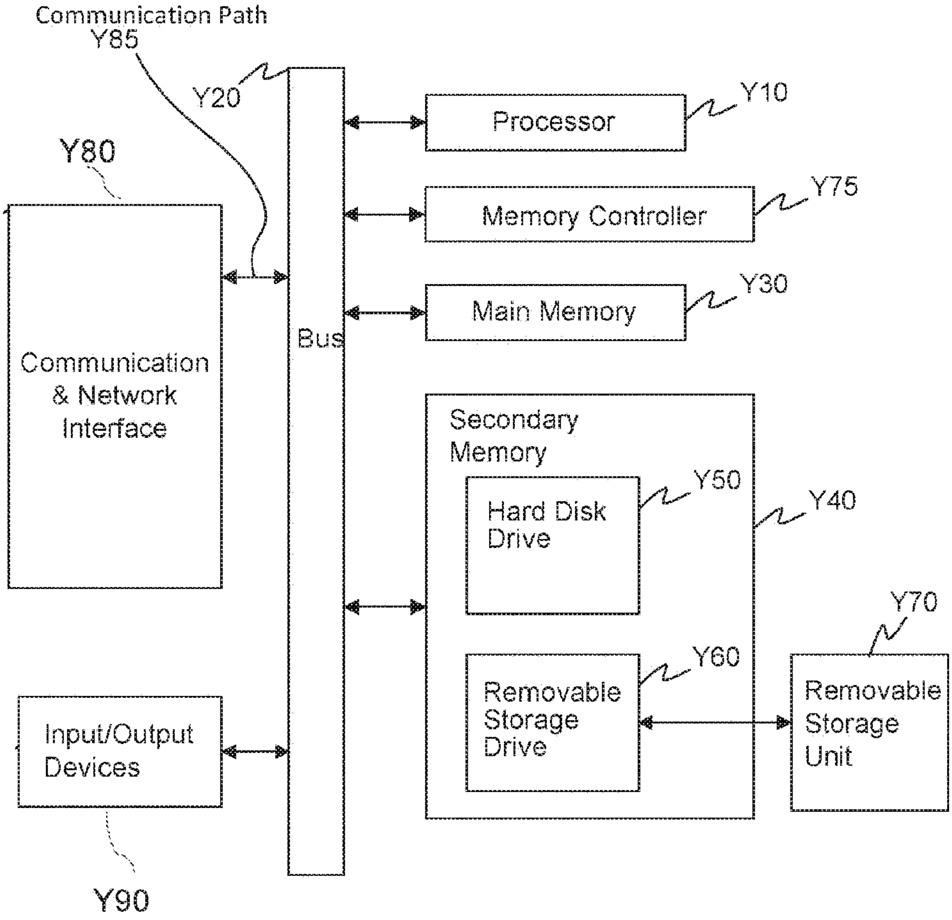

Communication Path
Y85
Y20
Y80
Communication & Network Interface
Bus
Processor
Y10
Memory Controller
Y75
Main Memory
Y30
Secondary Memory
Hard Disk Drive
Y50
Y40
Y70
Removable Storage Drive
Y60
Removable Storage Unit
Input/Output Devices
Y90

FIG. 6

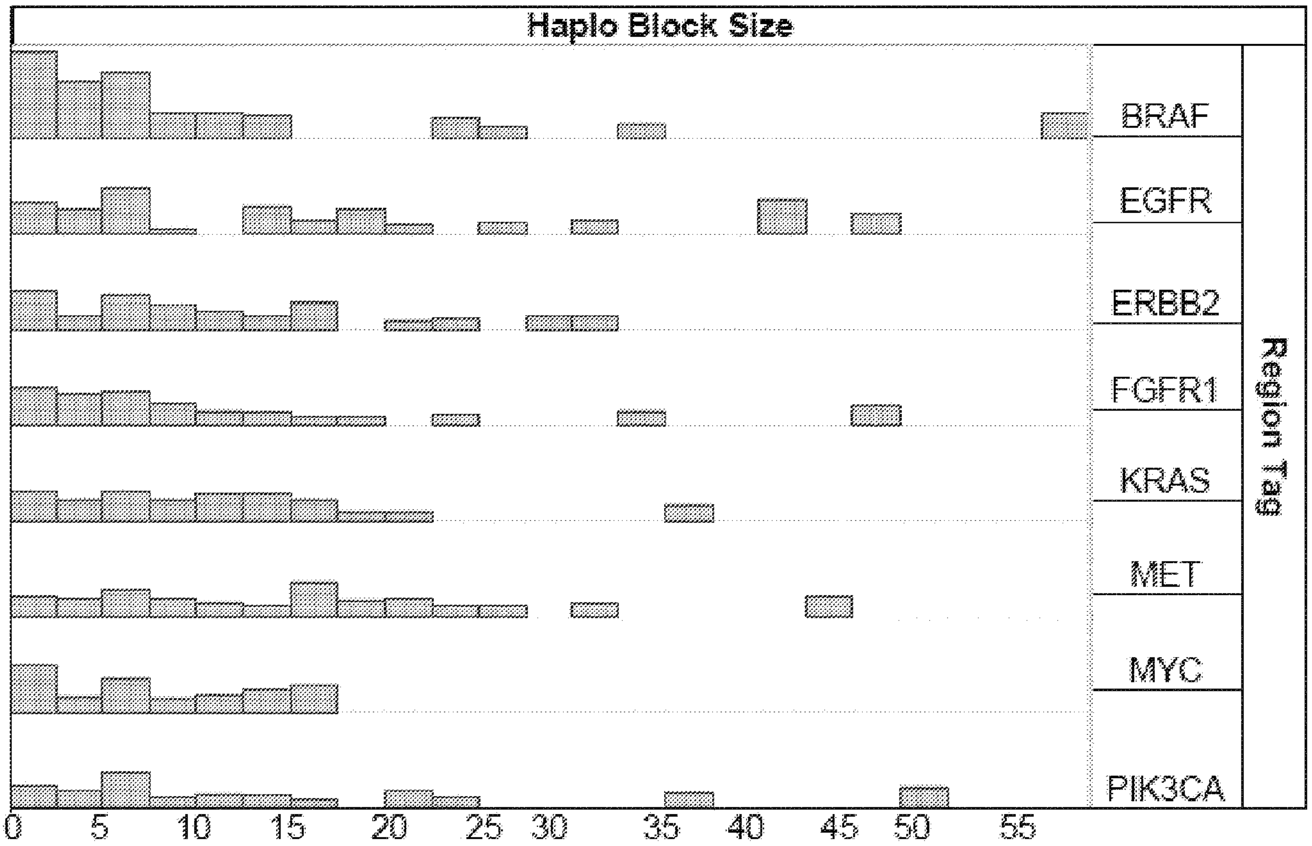
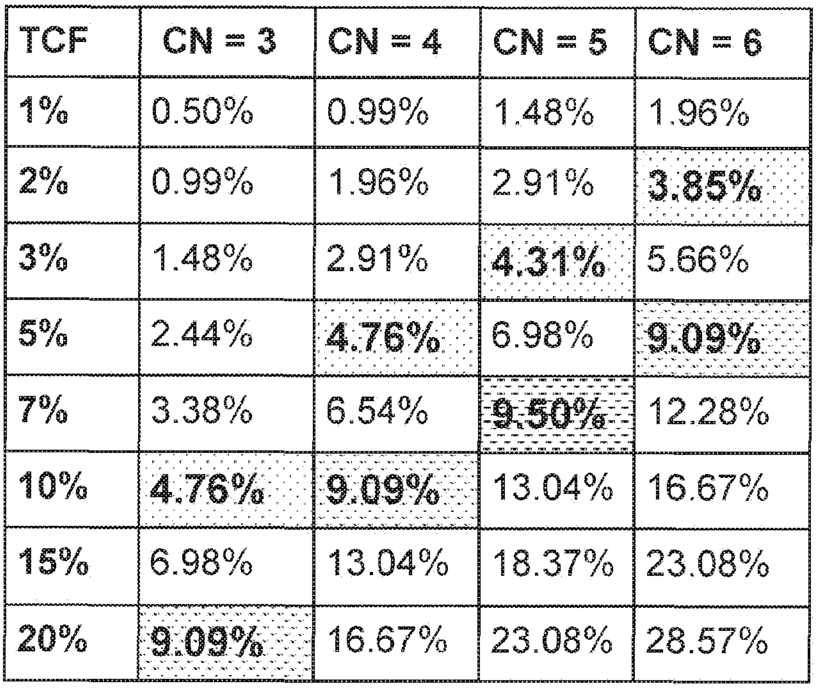

Haplo Block Size
Region Tag
BRAF
EGFR
ERBB2
FGFR1
KRAS
MET
MYC
PIK3CA
0
5
10
15
20
25
30
35
40
45
50
55

FIG. 7



TCF	CN = 3	CN = 4	CN = 5	CN = 6
1%	0.50%	0.99%	1.48%	1.96%
2%	0.99%	1.96%	2.91%	**3.85%**
3%	1.48%	2.91%	**4.31%**	5.66%
5%	2.44%	**4.76%**	6.98%	**9.09%**
7%	3.38%	6.54%	**9.50%**	12.28%
10%	**4.76%**	**9.09%**	13.04%	16.67%
15%	6.98%	13.04%	18.37%	23.08%
20%	**9.09%**	16.67%	23.08%	28.57%

FIG. 8