

USING SIZE AND NUMBER ABERRATIONS IN PLASMA DNA FOR DETECTING CANCER

CROSS-REFERENCES TO RELATED APPLICATIONS

[0001] This application claims priority to U.S. Patent Applications No. 62/102,867 entitled
5 “Using Size And Number Aberrations In Plasma DNA For Detecting Cancer” by Lo et al.
(attorney docket number 80015-015800US), filed January 13, 2015; and 62/111,534 entitled
“Using Size and Number Aberrations in Plasma DNA for Detecting Cancer” by Lo et al.
(attorney docket number 80015-015801US), filed February 3, 2015, the disclosures of which
are incorporated by reference in its entirety for all purposes.

BACKGROUND

[0002] The analysis of circulating cell-free DNA has been increasingly used for the
detection and monitoring of cancers (1-3). Different cancer-associated molecular
characteristics, including copy number aberrations (4-7), methylation changes (8-11), single
nucleotide mutations (4, 12-15), cancer-derived viral sequences (16, 17) and chromosomal
15 rearrangements (18, 19) can be detected in the plasma of patients with various types of
cancers. Despite the rapid expansion of clinical applications, many fundamental molecular
characteristics of circulating DNA in cancer patients remain unclear, thereby limiting the
most effective clinical use of such analyses.

[0003] In particular, previous studies on the size of circulating DNA in cancer patients gave
20 inconsistent results. Studies have demonstrated that the overall integrity (a measurement of
size) of circulating DNA would increase in cancer patients when compared with subjects
without a malignant condition (20-23). Using PCR with different amplicon sizes, it was
shown that the proportion of longer DNA would be higher in cancer patients. This aberration
in DNA integrity was shown to be reversible after treatment and the persistence of such
25 changes was associated with poor prognosis (20, 24). On the other hand, there is also
seemingly contradictory evidence that circulating DNA derived from tumor tissues might be
shorter than those derived from non-malignant cells. For example, it has been shown that the
proportion of DNA molecules carrying cancer-associated mutations would be higher when
those mutations were detected using PCR with shorter amplicons (12, 25).

[0004] Further, studying the size profile of tumor-derived DNA in the plasma of the HCC patients is a challenging endeavor because tumor-derived plasma DNA cannot be readily distinguished from the non-tumor-derived background DNA in plasma. The detection of cancer-specific mutations offers a genotypic means to distinguish the tumoral from the non-tumoral plasma DNA. However, there are relatively few cancer-specific mutations across the genome (29-32). Accordingly, it can be difficult to accurately identify tumor-derived DNA in plasma, particularly for the purpose of generating a broad, detailed and yet cost-effective view of the size distribution of tumor-derived DNA.

[0005] Such difficulties provide obstacles in obtaining accurate measurements in samples possibly containing mixtures of tumoral and non-tumoral DNA.

SUMMARY

[0006] Disclosed herein are systems and methods for determining whether regions exhibit an aberration (e.g., an amplification or a deletion), which may be associated with cancer. For example, examples can identify a region as possibly having an aberration using a count-based analysis and confirm whether the region does have the aberration using a size-based analysis.

[0007] In other examples, regions that exhibit an aberration can be compared to reference patterns that correspond to known types of cancer. A type of cancer can be identified when a sufficient number of regions have a matching aberration. Such matching regions can further be identified as related to the cancer for the analysis of tumor DNA, e.g., for a size analysis.

[0008] In yet other examples, a size analysis of DNA fragments in a sample (e.g., a mixture possibly containing both tumor and non-tumor DNA) can depend on a measured fraction of tumor DNA in the sample. For example, longer DNA fragments than healthy controls can indicate an early stage cancer for low tumor DNA fraction, and shorter DNA fragments than healthy controls can indicate a later stage cancer for higher tumor DNA fraction.

[0009] Other examples are directed to systems and computer readable media associated with methods described herein. In one aspect, the present invention refers to a method of analyzing a biological sample of an organism, the biological sample including nucleic acid molecules originating from normal cells and potentially from cells associated with cancer, wherein at least some of the nucleic acid molecules are cell-free in the biological sample, the method comprising: identifying a plurality of chromosomal regions of the organism, each chromosomal region including a plurality of loci; for each of a plurality of the nucleic acid molecules in the biological sample: measuring a size of the nucleic acid molecule; and identifying a location of the nucleic acid molecule in a reference genome of the organism; for each of the plurality of chromosomal regions: identifying a respective group of nucleic acid molecules as being from the

chromosomal region based on the identified locations, the respective group including at least one nucleic acid molecule located at each of the plurality of loci of the chromosomal region; calculating, with a computer system, a respective amount of the respective group of nucleic acid molecules; and comparing the respective amount to a count reference value to determine a count classification of whether the chromosomal region exhibits an aberration, the count classification for a first chromosomal region of the plurality of chromosomal regions indicating a first aberration; identifying a first group of nucleic acid molecules as being from the first chromosomal region based on the identified locations; calculating, with a computer system, a first statistical value of a first size distribution of the first group of nucleic acid molecules; and comparing the first statistical value to a size reference value to determine a size classification of whether the first chromosomal region exhibits the first aberration; and determining a final classification of whether the first chromosomal region exhibits the first aberration using the count classification and the size classification of the first chromosomal region. In another example, the present invention refers to a method of analyzing a biological sample of an organism, the biological sample including nucleic acid molecules originating from normal cells and potentially from cells associated with cancer, wherein at least some of the nucleic acid molecules are cell-free in the biological sample, the method comprising: for each of a plurality of the nucleic acid molecules in the biological sample: measuring a size of the nucleic acid molecule; identifying a location of the nucleic acid molecule in a reference genome of the organism; identifying a first group of nucleic acid molecules as being from a first chromosomal region based on the identified locations, the first chromosomal region including a plurality of first loci; calculating, with a computer system, a first statistical value of a size distribution of the first group of nucleic acid molecules, wherein the first statistical value is determined by computing an area under a first curve at a specified size, the first curve being a plot of cumulative frequency of nucleic acid molecules for the first chromosomal region over a range of sizes; and comparing the first statistical value to a first reference value to determine a classification of whether the first chromosomal region exhibits an aberration, wherein: the first reference value is a statistical value of a size distribution of a second group of nucleic acid molecules of a second chromosomal region, and the first reference value is determined by computing an area under a second curve at the specified size, the second curve being a plot of cumulative frequency of nucleic acid molecules for the second chromosomal region over the range of sizes. In a further aspect, the present invention refers to a method of analyzing a biological sample of an organism, the biological sample including nucleic acid molecules originating from normal cells and potentially from cells associated with cancer, wherein at least some of the nucleic acid molecules are cell-free in the biological sample, the method comprising: identifying a plurality of chromosomal regions of the organism, each chromosomal region including a plurality of loci, wherein the plurality of chromosomal regions are subchromosomal; for each of a plurality of nucleic acid molecules in the biological sample of

the organism: identifying a location of the nucleic acid molecule in a reference genome of the organism; for each of the plurality of chromosomal regions: identifying a respective group of nucleic acid molecules as being from the chromosomal region based on the identified locations, the respective group including at least one nucleic acid molecule located at each of the plurality of loci of the chromosomal region;

5 calculating, with a computer system, a respective value of the respective group of nucleic acid molecules, the respective value defining a property of the nucleic acid molecules of the respective group; and comparing the respective value to a respective reference value to determine a classification of whether the chromosomal region exhibits a deletion or an amplification; and determining a test pattern of the chromosomal regions that exhibit a deletion or amplification, the test pattern including: a set of the

10 chromosomal regions that exhibit a deletion or amplification; a first subset of the set that exhibit an amplification; and a second subset of the set that exhibit a deletion; comparing the test pattern to a plurality of reference patterns of different types of cancer; based on the comparison, determining a first amount of regions of the test pattern that exhibit a same deletion or amplification as a first reference pattern corresponding to a first type of cancer; and comparing the first amount to a first threshold to

15 determine a first classification of whether the biological sample exhibits the first type of cancer. In a further aspect, the present invention refers to a method of analyzing a biological sample of an organism, the biological sample including nucleic acid molecules originating from normal cells and from cells associated with cancer, wherein at least some of the nucleic acid molecules are cell-free in the biological sample, the method comprising: for each of a plurality of the nucleic acid molecules in the biological

20 sample: measuring a size of the nucleic acid molecule; identifying a location of the nucleic acid molecule in a reference genome of the organism; identifying a first group of nucleic acid molecules as being from a first chromosomal region based on the identified locations, the first chromosomal region including a plurality of first loci; calculating, with a computer system, a first statistical value of a size distribution of the first group of nucleic acid molecules; measuring a fraction of nucleic acid molecules originating from

25 cells associated with cancer; selecting a first reference value based on the measured fraction; comparing the first statistical value to the first reference value to determine a stage of cancer of the biological sample. In another aspect, the present invention refers to a computer product comprising a computer readable medium storing a plurality of instructions for controlling a computer system to perform an operation of any of the methods disclosed herein. In yet another aspect, the present invention refers to a

30 system comprising the computer product as disclosed herein; and one or more processors for executing instructions stored on the computer readable medium. In one aspect, the present invention refers to a system comprising means for performing any of the methods disclosed herein. In another aspect, the present invention refers to a system configured to perform any of the methods disclosed herein. In a

further aspect, the present invention refers to a system comprising modules that respectively perform the steps of any of the methods disclosed herein.

[0010] A better understanding of the nature and advantages of embodiments of the present invention may be gained with reference to the following detailed description and the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

5 **[0011]** FIG. 1 is a flowchart illustrating a method 100 of identifying chromosomal regions as exhibiting an aberration according to embodiments of the present invention.

[0012] FIG. 2 shows a Circos plot 200 identifying regions exhibiting amplifications and deletions in plasma and tissue samples of a representative hepatocellular carcinoma (HCC) patient according to embodiments of the present invention.

10 **[0013]** FIG. 3 shows plasma copy number aberration (CNA) results for various subjects according to embodiments of the present invention.

[0014] FIG. 4 is a table 400 showing detectability of CNA in plasma of HCC patients, hepatitis B virus (HBV) carriers, patients with liver cirrhosis and healthy subjects according to embodiments of the present invention.

15 **[0015]** FIG. 5 shows a table 500 of CNAs detected in the tumor and corresponding plasma of 12 HCC patients.

[0016] FIG. 6 shows a flowchart illustrating a method of analyzing a biological sample of an organism to determine whether a biological sample exhibits a first type of cancer according to embodiments of the present invention.

20 **[0017]** FIG. 7 shows chromosome arms that exhibit different patterns for different types of cancers in table 700 according to embodiments.

[0018] FIGs. 8A, 8B, and 8C show a table 800 of patterns of chromosomal regions for different types of cancer.

25 **[0019]** FIG. 9 shows a flowchart illustrating a method of analyzing a biological sample of an organism according to embodiments of the present invention.

[0020] FIG. 10 shows plots of the proportions of plasma DNA fragments of (A) shorter than 150 bp, (B) from 150 to 180 bp, and (C) longer than 180 bp against tumor DNA fraction in plasma.

[0021] FIG. 11 is a schematic illustration of the principle of plasma DNA size analysis in cancer patients.

[0022] FIG. 12 shows size distributions of plasma DNA originating from the amplified 8q and deleted 8p of a representative case H291. (A) The size distributions of plasma DNA for 8p (red) and 8q (green). (B) Plot of cumulative frequencies for plasma DNA size for 8p (red) and 8q (green). (C) The difference in cumulative frequencies for the HCC case H291.

[0023] FIG. 13 shows the difference in the cumulative frequencies for size between 8q and 8p (ΔS). (A) Plot of ΔS against size for all the HCC cases with different CNAs on 8p and 8q in plasma. (B) The values of ΔS_{166} amongst different groups.

[0024] FIG. 14 is a plot of the values of ΔS between 1q and 1p against size for a representative HCC patient.

[0025] FIG. 15 is a plot of the values of ΔS_{166} between 1q and 1p for healthy control subjects, HBV carriers, cirrhotic patients and HCC patients.

[0026] FIG. 16 is a flowchart illustrating a method of performing chromosome arm-level z-score analysis (CAZA) and size analysis in order to analyze a biological sample of an organism according to embodiments of the present invention.

[0027] FIG. 17 is a flowchart illustrating a method of analyzing a biological sample of an organism according to embodiments of the present invention.

[0028] FIG. 18 shows size distributions of plasma DNA fragments in the HCC patients with different fractional concentrations of tumor-derived DNA in plasma.

[0029] FIG. 19 shows size profiles of plasma DNA for (A) healthy controls, (B) chronic HBV carriers, and (C) cirrhotic patients.

[0030] FIG. 20 shows boxplots of the proportion of short fragments for healthy control subjects, HCC patients with tumor DNA fraction of less than 2% in plasma, and HCC patients with tumor DNA fraction of greater than 6%.

[0031] FIG. 21 is a receiver operating characteristic (ROC) curve for applying P(<150) to differentiate HCC patients with less than 2% tumor DNA fraction from healthy control subjects.

[0032] FIG. 22 is a receiver operating characteristic (ROC) curve for applying $P(<150)$ to differentiate HCC patients with greater than 6% tumor DNA fraction and healthy subjects.

[0033] FIG. 23 shows boxplots of the proportion of long fragments for healthy control subjects and HCC patients with tumor DNA fraction of less than 2% in plasma.

5 **[0034]** FIG. 24 is an ROC curve for using $P(>180)$ to differentiate HCC patients with less than 2% tumor DNA fraction from healthy control subjects.

[0035] FIG. 25 shows boxplots of median fragment size of healthy control subjects, HCC patients with less than 2% tumor DNA fraction, and HCC patients with greater than 6% tumor DNA fraction.

10 **[0036]** FIG. 26 is an ROC curve for using median fragment size to differentiate between HCC patients with less than 2% tumor DNA fraction and healthy control subjects.

[0037] FIG. 27 is an ROC curve for using median fragment size to differentiate between HCC patients with greater than 6% tumor DNA fraction and healthy control subjects.

15 **[0038]** FIG. 28 shows a boxplot of the proportion of short plasma DNA fragments of less than 150 bp that were aligned to chromosome 1q for HCC patients with greater than 6% tumor DNA fraction and for healthy control subjects.

[0039] FIG. 29 is an ROC curve for using the proportion of short plasma DNA fragments of less than 150 bp to differentiate between HCC patients with greater than 6% tumor DNA fraction and healthy control subjects.

20 **[0040]** FIG. 30 is a plot of ΔS versus tumor size of HCC patients.

[0041] FIG. 31 is a plot of the percentage of DNA fragments of a certain size against tumor size.

[0042] FIG. 32 shows a block diagram of an example computer system 10 usable with system and methods according to embodiments of the present invention.

25

TERMS

[0043] The term “*biological sample*” as used herein refers to any sample that is taken from a subject (*e.g.*, a human, such as a pregnant woman) and contains one or more nucleic acid molecule(s) of interest. Examples include plasma, saliva, pleural fluid, sweat, ascitic fluid, bile, urine, serum, pancreatic juice, stool, cervical lavage fluid, and cervical smear samples.

[0044] The term “*nucleic acid*” or “*polynucleotide*” refers to a deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) and a polymer thereof in either single- or double-stranded form. Unless specifically limited, the term encompasses nucleic acids containing known analogs of natural nucleotides that have similar binding properties as the reference nucleic acid and are metabolized in a manner similar to naturally occurring nucleotides. Unless otherwise indicated, a particular nucleic acid sequence also implicitly encompasses conservatively modified variants thereof (*e.g.*, degenerate codon substitutions), alleles, orthologs, single nucleotide polymorphisms (SNPs), and complementary sequences as well as the sequence explicitly indicated. Specifically, degenerate codon substitutions may be achieved by generating sequences in which the third position of one or more selected (or all) codons is substituted with mixed-base and/or deoxyinosine residues (Batzer MA *et al.*, *Nucleic Acids Res* 1991; **19**:5081; Ohtsuka E *et al.*, *J Biol Chem* 1985; **260**:2605-2608; and Rossolini GM *et al.*, *Mol Cell Probes* 1994; **8**:91-98). The term nucleic acid is used interchangeably with gene, cDNA, mRNA, small noncoding RNA, microRNA (miRNA), Piwi-interacting RNA, and short hairpin RNA (shRNA) encoded by a gene or locus.

[0045] The term “*gene*” means the segment of DNA involved in producing a polypeptide chain. It may include regions preceding and following the coding region (leader and trailer) as well as intervening sequences (introns) between individual coding segments (exons).

[0046] As used herein, the term “*locus*” or its plural form “*loci*” is a location or address of any length of nucleotides (or base pairs) which has a variation across genomes.

[0047] The term “*sequenced tag*” (also called sequence read) refers to a sequence obtained from all or part of a nucleic acid molecule, *e.g.*, a DNA fragment. In one embodiment, just one end of the fragment is sequenced, *e.g.*, about 30 bp. The sequenced tag can then be aligned to a reference genome. Alternatively, both ends of the fragment can be sequenced to generate two sequenced tags, which can provide greater accuracy in the alignment and also provide a length of the fragment. In yet another embodiment, a linear DNA fragment can be circularized, *e.g.*, by ligation, and the part spanning the ligation site can be sequenced.

[0048] The term fractional tumor DNA concentration is used interchangeably with the terms tumor DNA proportion and tumor DNA fraction, and refers to the proportion of DNA molecules that are present in a sample that is derived from a tumor.

[0049] The term “*size profile*” generally relates to the sizes of DNA fragments in a biological sample. A size profile may be a histogram that provides a distribution of an

amount of DNA fragments at a variety of sizes. Various statistical parameters (also referred to as size parameters or just parameter) can be used to distinguish one size profile to another. One parameter is the percentage of DNA fragment of a particular size or range of sizes relative to all DNA fragments or relative to DNA fragments of another size or range.

5 **[0050]** The term “*parameter*” as used herein means a numerical value that characterizes a quantitative data set and/or a numerical relationship between quantitative data sets. For example, a ratio (or function of a ratio) between a first amount of a first nucleic acid sequence and a second amount of a second nucleic acid sequence is a parameter.

10 **[0051]** The term “*classification*” as used herein refers to any number(s) or other characters(s) (including words) that are associated with a particular property of a sample. For example, a “+” symbol could signify that a sample is classified as having deletions or amplifications (e.g., duplications). The terms “*cutoff*” and “*threshold*” refer to a predetermined number used in an operation. For example, a cutoff size can refer to a size above which fragments are excluded. A threshold value may be a value above or below
15 which a particular classification applies. Either of these terms can be used in either of these contexts.

20 **[0052]** The term “*level of cancer*” can refer to whether cancer exists, a stage of a cancer, a size of tumor, how many deletions or amplifications of a chromosomal region are involved (e.g. duplicated or triplicated), and/or other measure of a severity of a cancer. The level of cancer could be a number or other characters. The level could be zero. The level of cancer also includes premalignant or precancerous conditions associated with deletions or amplifications.

25 **[0053]** A “subchromosomal region” is a region that is smaller than a chromosome. Examples of subchromosomal regions are 100 kb, 200 kb, 500 kb, 1 Mb, 2 Mb, 5 Mb, or 10 Mb. Another example of a subchromosomal region is one that corresponds to one or more
bands, or subbands, or one of the arms of a chromosome. Bands or subbands are features observed in cytogenetic analysis. A subchromosomal region may be referred to by its genomic coordinates in relation to a reference human genome sequence.

DETAILED DESCRIPTION

30 **[0054]** Cancers often have regions with copy number aberrations (amplifications or deletions) relative to the person’s normal genome. Techniques can count cell-free DNA fragments in a sample (e.g., plasma or serum) that include tumor DNA fragment and non-

tumor DNA fragments. The counting can identify regions that are over-represented (indicative of amplification) or under-represented (indicative of deletion). But, as such count-based techniques are statistical in nature, incorrect indications can occur. Embodiments can identify a region as possibly having a copy number aberration (also referred to as aberration) using a count-based analysis and confirm whether the region does have the aberration using a size-based analysis. Such a confirmation provides additional accuracy in identifying regions with aberrations.

[0055] Regions that have aberrations can be used to identify an existence of cancer in the organism from which the sample was obtained. But, the existence of cancer does not convey a type of cancer. To address this problem, embodiments can use reference patterns of aberrations in regions from samples with known cancers. A test pattern of which regions are aberrant can be determined for a given sample being tested, and the test pattern can be compared to the references patterns to determine a type of cancer. An amount of regions of the test pattern that exhibit a same deletion or amplification as a reference pattern corresponding to a particular type of cancer can be determined, and the amount can be compared to a threshold to determine a classification of whether the particular type of cancer is present. Once a region is identified as both having an aberration and corresponding to a particular type of cancer, one can have greater confidence in analyzing the region for tumor DNA. For example, the region can be used to measure a tumor DNA fraction in the sample.

[0056] Additionally, various studies have shown inconsistent results as to the length of cell-free tumor DNA fragments: some showing longer fragments for tumor DNA and other showing shorter fragments for tumor DNA. The analysis below shows that both can be correct, but for different tumor DNA fractions. Embodiments can use different size thresholds in a size-analysis based on a measured tumor DNA fraction, which may be determined using counting of DNA fragments in a region identified as having an aberration. Accordingly, some implementations can reconcile these apparent inconsistencies through, for example: (a) genome-wide high resolution size profiling of plasma DNA enabled by massively parallel sequencing; and (b) an efficient approach to distinguish tumor-derived DNA from the non-tumoral background DNA in the plasma of cancer patients (e.g., using regions identified as having an aberration).

I. INTRODUCTION

[0057] It has become feasible to measure the lengths of millions or billions of every individual plasma DNA molecule in samples with the use of massively parallel sequencing (26, 27). Hence, plasma DNA sizes could be studied in a genomewide manner and at single-
5 base resolution. Using this approach, the size of circulating DNA has generally been shown to resemble the size of mononucleosomal DNA suggesting that plasma DNA might be generated through apoptosis (26, 27). In pregnant women, plasma DNA derived from the fetus has been shown to be shorter than that of DNA derived from the mother (26). The size difference between circulating fetal and maternal DNA has provided a new conceptual basis
10 for quantifying fetal DNA in maternal plasma and detecting chromosomal aneuploidies through size analysis of plasma DNA (28). In addition, differences in the size distributions of circulating DNA derived from the transplanted organs and the patients' own tissues have been observed for recipients of solid organ or bone marrow transplantation (27).

[0058] Plasma of cancer patients contains a mixture of tumor-derived DNA and non-tumor-
15 derived DNA. Examples below analyze the size distribution of plasma DNA in cancer patients with hepatocellular carcinoma (HCC). The size distributions of plasma DNA in HCC patients, patients with chronic hepatitis B virus (HBV) infection, patients with liver cirrhosis and healthy subjects were also analyzed. Embodiments can use certain aberrant regions to analyze the size profile of tumor-derived DNA in the plasma of the HCC patients. The use of
20 such aberrant regions can overcome the challenge that tumor-derived plasma DNA is not readily distinguished from the non-tumor-derived background DNA in plasma.

[0059] Some embodiments use chromosome arms that are affected by copy number aberrations (CNAs) to infer the difference in size distributions of tumor- and non-tumor-derived plasma DNA. For chromosome arms that are amplified in the tumor tissues, the
25 proportional contribution from tumor-derived DNA to plasma DNA would increase whereas for chromosome arms that are deleted in the tumor, the contribution would decrease. Therefore, the comparison of size profiles of chromosome arms that are amplified and deleted would reflect the size difference between tumor-derived and non-tumor-derived DNA in plasma. CNAs involving a whole chromosome arm or a large trunk of a chromosome arm is
30 relatively common (33). Deletion of chromosomes 1p and 8p and amplification of chromosomes 1q and 8q are commonly observed in the HCC tissues (34-36). Thus, the

analysis focuses on chromosomes 1 and 8 for the CNA and size profiling analyses of plasma DNA.

II. COUNTING ANALYSIS TO IDENTIFY ABERRANT REGIONS

[0060] An aberrant region includes an amplification or a deletion. An amplification means
5 that a sequence in the region occurs more often than it does in a reference sequence, and thus
the sequence has been amplified. The amplification typically would occur in only one
chromosome copy (haplotype). A deletion means that a sequence in the region has been
deleted relative to the reference sequence, typically just one chromosome copy has the
deletion for diploid organisms. A region can be defined by at least two loci (which are
10 separated from each other), and DNA fragments at these loci can be used to obtain a
collective value about the region.

A. Detecting an Aberrant Region by Counting

[0061] The aberration of a region can be determined by counting an amount of DNA
fragments (molecules) that are derived from the region. As examples, the amount can be a
15 number of DNA fragments, a number of bases to which a DNA fragment overlapped, or other
measure of DNA fragments in a region. The amount of DNA fragments for the region can be
determined by sequencing the DNA fragments to obtain sequence reads and aligning the
sequence reads to a reference genome. In one embodiment, the amount of sequence reads for
the region can be compared to the amount of sequence reads for another region so as to
20 determine overrepresentation (amplification) or underrepresentation (deletion). In another
embodiment, the amount of sequence reads can be determined for one haplotype and
compared to the amount of sequence reads for another haplotype.

[0062] Accordingly, the number of DNA fragments from one chromosomal region (e.g., as
determined by counting the sequenced tags aligned to that region) can be compared to a
25 reference value (which may be determined from a reference chromosome region, from the
region on another haplotype, or from the same region in another sample that is known to be
healthy). The comparison can determine whether the amount is statistically different (e.g.,
above or below) the reference value. A threshold for the difference can be used, e.g.,
corresponding to 3 standard deviations (SD), as seen in a distribution of values seen in a
30 population.

[0063] As part of the comparison, a tag count can be normalized before the comparison. A normalized value for the sequence reads (tags) for a particular region can be calculated by dividing the number of sequenced reads aligning to that region by the total number of sequenced reads alignable to the whole genome. This normalized tag count allows results from one sample to be compared to the results of another sample. For example, the normalized value can be the proportion (e.g., percentage or fraction) of sequence reads expected to be from the particular region. But, many other normalizations are possible, as would be apparent to one skilled in the art. For example, one can normalize by dividing the number of counts for one region by the number of counts for a reference region (in the case above, the reference region is just the whole genome) or by always using a same number of sequence reads. This normalized tag count can then be compared against a threshold value, which may be determined from one or more reference samples not exhibiting cancer.

[0064] In some embodiments, the threshold value can be the reference value. In other embodiments, the reference value can be the other value used for normalization, and the comparison can include the reference value and the threshold value. For example, the amount for the region can be divided by the reference value to obtain a parameter, which is compared to the threshold value to see if a statistically significant difference exists. As another example, the amount for the region can be compared to the reference value plus the threshold value.

[0065] In one embodiment, the comparison is made by calculating the z-score of the case for the particular chromosomal region. The z-score can be calculated using the following equation: $z\text{-score} = (\text{normalized tag count of the case} - \text{mean}) / \text{SD}$, where “mean” is the mean normalized tag count aligning to the particular chromosomal region for the reference samples; and SD is the standard deviation of the number of normalized tag count aligning to the particular region for the reference samples. Hence, the z-score can correspond to the number of standard deviations that the normalized tag count of a chromosomal region for the tested case is away from the mean normalized tag count for the same chromosomal region of the one or more reference subjects. This z-score can be compared to a threshold, e.g., 3 for amplification and -3 for deletion. Chromosomal regions that are amplified would have a positive value of the z-score above the threshold. Chromosomal regions that are deleted would have a negative value of the z-score that is below the threshold.

[0066] The magnitude of the z-score can be determined by several factors. One factor is the fractional concentration of tumor-derived DNA in the biological sample (e.g. plasma). The

higher the fractional concentration of tumor-derived DNA in the sample (e.g. plasma), the larger the difference between the normalized tag count of the tested case and the reference cases would be. Hence, a larger magnitude of the z-score would result.

[0067] Another factor is the variation of the normalized tag count in the one or more reference cases. With the same degree of the over-representation of the chromosomal region in the biological sample (e.g. plasma) of the tested case, a smaller variation (i.e. a smaller standard deviation) of the normalized tag count in the reference group would result in a higher z-score. Similarly, with the same degree of under-representation of the chromosomal region in the biological sample (e.g. plasma) of the tested case, a smaller standard deviation of the normalized tag count in the reference group would result in a more negative z-score.

[0068] Another factor is the magnitude of chromosomal aberration in the tumor tissues. The magnitude of chromosomal aberration refers to the copy number changes for the particular chromosomal region (either gain or loss). The higher the copy number changes in the tumor tissues, the higher the degree of over- or under-representation of the particular chromosomal region in the plasma DNA would be. For example, the loss of both copies of the chromosome would result in greater under-representation of the chromosomal region in the plasma DNA than the loss of one of the two copies of the chromosome and, hence, resulted in a more negative z-score. Typically, there are multiple chromosomal aberrations in cancers. The chromosomal aberrations in each cancer can further vary by its nature (i.e. amplification or deletion), its degree (single or multiple copy gain or loss) and its extent (size of the aberration in terms of chromosomal length).

[0069] The precision of measuring the normalized tag count is affected by the number of molecules analyzed. For example, 15,000, 60,000 and 240,000 molecules may be needed to be analyzed to detect chromosomal aberrations with one copy change (either gain or loss) when the fractional concentration is approximately 12.5%, 6.3% and 3.2% respectively. Further details of the tag counting for detection of cancer for different chromosomal regions is described in U.S. Patent Publication No. 2009/0029377 entitled “Diagnosing Fetal Chromosomal Aneuploidy Using Massively Parallel Genomic Sequencing” by Lo et al; and U.S. Patent No. 8,741,811 entitled “Detection Of Genetic Or Molecular Aberrations Associated With Cancer” by Lo et al., the disclosure of which are incorporated by reference in its entirety for all purposes.

B. *Method*

[0070] FIG. 1 is a flowchart illustrating a method 100 of identifying a chromosomal region as potentially exhibiting an amplification according to embodiments of the present invention. Method 100, and other methods described herein, can be performed entirely or partially using
5 a computer system.

[0071] At step 110, a plurality of chromosomal regions of an organism may be identified. Each chromosomal region may include a plurality of loci. A region may be 1 Mb in size, or some other equal size. The entire genome can then include about 3,000 regions, each of predetermined size and location. Such predetermined regions can vary to accommodate a
10 length of a particular chromosome or a specified number of regions to be used, and any other criteria mentioned herein. If regions have different lengths, such lengths can be used to normalize results, e.g., as described herein.

[0072] Steps 120-140 may be performed for each of the chromosomal regions. At step 120, for each chromosomal region, a respective group of nucleic acid molecules may be identified
15 as being from the chromosomal region. The identification may be based on identifying a location of nucleic acid molecules in a reference genome. For example, the cell-free DNA fragments can be sequenced to obtain sequence reads, and the sequence reads can be mapped (aligned) to the reference genome. If the organism was a human, then the reference genome would be a reference human genome, potentially from a particular subpopulation. As another
20 example, the cell-free DNA fragments can be analyzed with different probes (e.g., following PCR or other amplification), where each probe corresponds to a different genomic location. In some embodiments, the analysis of the cell-free DNA fragments can be performed by receiving sequence reads or other experimental data corresponding to the cell-free DNA fragments, and then analyzing the experimental data using a computer system.

[0073] At step 130, a computer system may calculate a respective amount of the respective
25 group of nucleic acid molecules. The respective value defines a property of the nucleic acid molecules of the respective group. The respective value can be any of the values mentioned herein. For example, the value can be the number of fragments in the group or a statistical value of a size distribution of the fragments in the group. The respective value can also be a
30 normalized value, e.g., a tag count of the region divided by the total number of tag counts for the sample or the number of tag counts for a reference region. The respective value can also

be a difference or ratio from another value, thereby providing the property of a difference for the region.

[0074] At step 140, the respective amount may be compared to a reference value to determine a classification of whether the chromosomal region exhibits an aberration (i.e. an amplification or a deletion). In some embodiments, the chromosomal region may be classified as not exhibiting an aberration. The comparison may include determining a z-score based on the respective amount and the reference value. As an example, the reference value may be any threshold or reference value described herein. For example, the reference value could be a threshold value determined for normal samples. As another example, the reference value could be the tag count for another region, and the comparison can include taking a difference or ratio (or function of such) and then determining if the difference or ratio is greater than a threshold value.

[0075] The reference value may vary based on the results of other regions. For example, if neighboring regions also show a deviation (although small compared to a threshold, e.g., a z-score of 3), then a lower threshold can be used. For example, if three consecutive regions are all above a first threshold, then cancer may be more likely. Thus, this first threshold may be lower than another threshold that is required to identify cancer from non-consecutive regions. Having three regions (or more than three) having even a small deviation can have a low enough probability of a chance effect that the sensitivity and specificity can be preserved.

C. Chromosome Arm-level Z-score Analysis (CAZA)

[0076] In some embodiments, a chromosome can be split into many subchromosomal regions (e.g., 1 Mb regions). This high resolution may not maximize sensitivity and specificity. Other embodiments can split a chromosome into two arms, namely p and q. Analyzing the two arms can improve specificity by reducing noise caused by such fine resolution. An example of chromosome arm-level z-score analysis is now provided.

[0077] We analyzed a total of 225 plasma DNA samples from 90 HCC patients, 67 patients with chronic HBV infection, 36 patients with HBV-associated liver cirrhosis and 32 healthy subjects. A median of 31 million reads (range: 17-79 million) was obtained from each plasma sample. Amounts of sequence reads originating from chromosome arms that were three SDs below (z-scores < -3) and three SDs above (z-scores > 3) the mean of healthy controls were deemed to indicate significant under- and over-representations of the plasma DNA from those

chromosome arms, respectively. These quantitative plasma DNA aberrations were generally reflective of the presence of copy number losses and copy number gains (CNAs) in the tumor (4).

[0078] FIG. 2 shows a Circos plot 200 identifying regions exhibiting amplifications and deletions in plasma and tissue samples of a representative hepatocellular carcinoma (HCC) patient according to embodiments of the present invention. From inside to outside: CNAs in the tumor tissue (in 1-Mb resolution); arm-level CNAs in the tumor tissue; plasma CNAs (in 1-Mb resolution); arm-level plasma CNAs. Regions with gains and losses are shown in green and red, respectively. The distance between two consecutive horizontal lines represents a z-score of 5. Chromosome ideograms (outside the plots) are oriented from pter to qter in a clockwise direction.

[0079] FIG. 3 shows plasma copy number aberration (CNA) results for all the studied subjects using an embodiment of CAZA. The four chromosome arms (1p, 1q, 8p and 8q) that are frequently affected by CNAs in HCC were analyzed. Red and green lines represent under- and over-representation, respectively, of the corresponding chromosome arms in plasma. Each vertical line represents the data for one case.

[0080] FIG. 4 is a table 400 showing detectability of CNA in plasma of HCC patients, HBV carriers, patients with liver cirrhosis and healthy subjects according to embodiments of the present invention. Table 400 shows categories of patients in the leftmost column. The remaining columns show the number of patients and the percentage with CNA detected in the plasma for different chromosome arms. Seventy-six (84.4%) of the 90 HCC patients had at least one chromosomal arm-level CNA on chromosomes 1 and 8 in plasma. Tumor tissues of 12 HCC patients were available to corroborate the plasma DNA findings. The tissue samples were sequenced and the CNA patterns are shown in FIG. 5.

[0081] FIG. 5 shows a table 500 of CNAs detected in the tumor and corresponding plasma of 12 HCC patients. In table 500, the patient case number is listed in the first column. The patients are arranged in descending order of tumor DNA fraction in plasma, as shown in the second column. The third column shows the tumor size. The remaining columns show CNAs detected in the tumor and plasma for different chromosome arms. 'Gain' indicates a copy number gain. 'Loss' indicates a copy number loss. 'Nil' indicates no detectable CNA. A total of 48 chromosome arms were analyzed for the 12 patients. The numbers (and percentages) of

chromosome arms with concordant and discordant results between tumor and plasma are shown.

[0082] Of the 48 chromosome arms analyzed for the 12 patients, concordant changes in plasma and tumor tissues were observed for 30 (63%) arms. CNAs were only observed in the tumor, but not in the plasma, for 10 (21%) arms. These cases tended to have lower tumor DNA fractions in plasma. CNAs were observed in the plasma, but not the tumor, for 7 (15%) arms. In one case (HOT428), a gain of 1q was observed in the tumor, but a loss was observed in plasma. These data might suggest the presence of tumoral heterogeneity where there might be other foci or clones of cancer cells contributing plasma DNA.

[0083] Among the HBV carriers with and without liver cirrhosis, the detection rates of these CNA were 22.2% and 4.5%, respectively. One patient with liver cirrhosis and one chronic HBV carrier without cirrhosis exhibited CNAs in plasma, but not known to have HCC at the time of blood collection, were diagnosed as having HCC at 3 months and 4 months afterwards, respectively. All the HBV carriers and cirrhotic patients were followed up for at least 6 months. For those control subjects without any CNA in plasma, none of them had developed HCC during the follow-up period. None of the 32 healthy subjects had detectable CNA on chromosome 1 or 8 in plasma by CAZA. In the HCC patients, the disproportionate increase or decrease in sequence reads in plasma due to the presence of CNA is reflective of the fractional concentration of tumor DNA in the plasma sample. The median fractional concentration of tumor-derived DNA in the plasma of the HCC patients was 2.1% (range: 0% to 53.1%; interquartile range: 1.2% to 3.8%).

[0084] CAZA provides a way to detect tumor-associated CNAs non-invasively. In HCC, chromosomes 1 and 8 are commonly affected by CNAs (34-36). Indeed, our data showed that 76 (84.4%) of the 90 HCC patients had at least one CNA involving either arms on chromosomes 1 and 8 in plasma, whereas none of the 32 healthy subjects exhibited any CNA for these two chromosomes in plasma. Plasma CNAs involving chromosomes 1 and 8 were also detected in 22.2% and 4.5% of the cirrhotic patients and HBV carriers. In one HBV carrier and one patient with liver cirrhosis, HCC was diagnosed shortly after the blood collection. It is likely that the cancer would have been present at the time of blood collection and was associated with the CNAs in plasma, thereby showing the early screening capabilities of embodiments. The relatively high detection rate of plasma CNAs in the HCC patients suggests that this approach might have future value in the screening of HBV carriers.

Moreover, CNAs are present in almost all types of cancer (33). Therefore, this approach can be applied as a generic tumor marker with adaptation to the specific CNA patterns of the cancer of interest.

III. DETECTING CANCER TYPE BASED ON PATTERN OF ABERRANT REGIONS

[0085] Some embodiments can use known aberrant regions (along with whether amplification or deletion) of a type of cancer in order to identify potential cancers implicated by aberrations identified in the sample. In the example above, the known aberrant regions for HCC were used to screen the sample for HCC. This screening can compare the identified aberrant regions (including whether amplification or deletion) to a known set of aberrant regions. If a sufficiently high match is determined, then that type of cancer can be flagged as a possible test result.

[0086] A matching criteria can be the percentage of regions of the set that are also identified in the sample. The matching criteria can require specific regions to be aberrant. For example, the match can be identified for HCC when 1p, 1q, or 8q is aberrant, or when more than one of these chromosome arms are aberrant. Thus, there can be specific subsets to which identical match is required, but the subsets can be smaller than a full set of known aberrant regions for a type of cancer.

[0087] Thus, a pattern of aberrant regions for a test sample can be compared to the pattern of aberrant regions for a particular type of cancer, which may be determined from patients known to have a particular type of cancer. Embodiments can be used to screen for cancer and identify the type of cancer involved, particularly where the tumor may be small (e.g., less than 2 cm in size). Imaging techniques have difficulty in identifying tumors less than 2 cm in size. Such techniques can also be used to track progress of the patient after treatment.

A. Method

[0088] FIG. 6 is a flowchart illustrating a method 600 of analyzing a biological sample of an organism to determine whether a biological sample exhibits a first type of cancer according to embodiments of the present invention. The biological sample includes nucleic acid molecules (also called fragments) originating from normal cells and potentially from cells associated with cancer. At least some of these molecules may be cell-free in the sample.

[0089] In one embodiment of this and any other method described herein, the biological sample includes cell-free DNA fragments. Although the analysis of plasma DNA has been used to illustrate the different methods described in this application, these methods can also be applied to detect tumor-associated chromosomal aberrations in samples containing a mixture of normal and tumor-derived DNA. The other sample types include saliva, tears, pleural fluid, ascitic fluid, bile, urine, serum, pancreatic juice, stool and cervical smear samples

[0090] In step 610, a plurality of chromosomal regions of the organism are identified. The plurality of chromosomal regions are subchromosomal and may be non-overlapping. The chromosomal regions that are counted can have restrictions. For example, only regions that are contiguous with at least one other region may be counted (or contiguous regions can be required to be of a certain size, e.g., four or more regions). For embodiments where the regions are not equal, the number can also account for the respective lengths (e.g., the number could be a total length of the aberrant regions). In some embodiments, the regions correspond to arms of the chromosomes. In other embodiments, the regions may be smaller than the arms, e.g., 1-Mb regions.

[0091] In some embodiments, a chromosomal region can be of a particular haplotype (i.e., correspond to a particular chromosome copy). In embodiments using a relative haplotype dosage (RHDO) analysis, each region can include at least two heterozygous loci. Further details on RHDO can be found in U.S. Patent No. 8,741,811.

[0092] In step 620, for each of a plurality of nucleic acid molecules in the biological sample of the organism, a location of the nucleic acid molecule in a reference genome of the organism can be identified. The plurality of nucleic acid molecules may include 500,000 or more molecules (fragments). This locating can be performed in various ways, including performing a sequencing of a molecule (e.g. via a random sequencing), to obtain one or two (paired-end) sequenced tags of the molecule and then aligning the sequenced tag(s) to the reference genome. Such alignment can be performed using such as tools as basic local alignment search tool (BLAST). The location can be identified as a number in an arm of a chromosome.

[0093] In step 630, a respective group of nucleic acid molecules may be identified as being from the chromosomal region based on the identified region, for each of the plurality of

chromosomal regions. The respective group may include at least one nucleic acid molecule located at each of the plurality of loci of the chromosomal region.

[0094] In step 640, a computer system may calculate a respective value of the respective group of nucleic acid molecules for each of the plurality of chromosomal regions. The respective value may define a property of the nucleic acid molecules of the respective group. The property may be a count, a percentage, or a size of the nucleic acid molecules. The respective value may include a mean of a size distribution, a median of the size distribution, a mode of the size distribution, or a proportion of nucleic acid molecules having a size below a size threshold. Using size as a property is discussed in greater detail in Section IV.

[0095] In step 650, the respective value may be compared to a respective reference value to determine a classification of whether the chromosomal region exhibits a deletion or an amplification. The comparison may include determining a z-score based on the respective value and the respective reference value. The z-score can then be compared to one or more threshold values to determine whether a deletion or an amplification exists. Different thresholds can be used for a deletion and an amplification. In other embodiments, the reference value can include the threshold value, e.g., if the other values in the z-score were moved to the other side of the equation. A reference value can correspond to a value determined in a healthy sample, another chromosomal region (e.g., one not exhibiting an aberration), or the other haplotype when the region being tested is a first haplotype.

[0096] In step 660, a test pattern of the chromosomal regions that exhibit a deletion or amplification may be determined. The test pattern refers to the pattern of aberrant regions in the sample being tested. The test pattern may include a set of chromosomal regions that exhibit a deletion, an amplification, or are normal. The test pattern may also include a first subset of the set that is identified as exhibiting an amplification. The test pattern may further include a second subset of the set that is identified as exhibiting a deletion. The test pattern can further include a third subset of the set that is identified as not exhibiting an amplification or a deletion.

[0097] In step 670, the test pattern may be compared to a plurality of reference patterns of different types of cancer. A reference patterns for a type of cancer may include a known set of aberrant regions. The reference patterns may be determined from reference samples of tissues and/or mixtures of cell-free nucleic acid molecules. The reference pattern may include a number of regions, with each having a defined status of amplification, deletion, or no

aberration. The comparison can determine which regions of the test pattern have a same aberration as regions in a reference pattern. For example, it can be determined whether the same region is indicated as having an amplification, a deletion, or is normal in both the test pattern and a reference pattern.

5 **[0098]** In step 680, based on the comparison, an amount of regions of the test pattern that exhibit a same deletion or amplification as a first reference pattern corresponding to a first type of cancer can be determined. In various embodiments, the amount may be a number or percentage of chromosomal regions that match with the known set of aberrant regions.

10 **[0099]** In step 690, the amount of regions is compared to a first threshold to determine a first classification of whether the biological sample exhibits the first type of cancer. The first threshold may be specific to the first type of cancer or be used across multiple types of cancer. Such a threshold may be a minimum amount of chromosomal regions needed to match with the known set of aberrant regions for the first type of cancer to be identified. In various
15 embodiments, the minimum amount may be 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, or 13 chromosomal regions. In some embodiments, specific regions may be required to be aberrant, and thus other criteria can be used besides the comparison of the amount to the first threshold. Such specific regions can be a constraint or be weighted higher than other regions. The specific aberrant regions may be a subset of the full set of known aberrant regions for a type of cancer. The type of cancer may include HCC, colorectal cancer, breast cancer, lung cancer, or
20 nasopharyngeal carcinoma, among other cancers.

25 **[0100]** A threshold value used to determine the classification may vary based on the locations and the sizes of the regions that are counted. For example, the amount of regions on a particular chromosome or arm of a chromosome may be compared to a threshold for that particular chromosome (or arm) as a criterion for determining whether a particular type of cancer is implicated. Multiple thresholds may be used. For instance, the amount of matching regions (i.e., same classification of aberration in test pattern and reference pattern) on a particular chromosome (or arm or larger subchromosomal region) may be required to be greater than a first threshold value, and the total amount of matching regions in the genome may be required to be greater than a second threshold value.

30 **[0101]** The threshold value for the amount of matching regions can also depend on how strong the imbalance is for the classification of the regions. For example, the amount of matching regions that are used as the threshold for determining a classification of a type of

cancer can depend on the specificity and sensitivity (aberrant threshold) used to detect an aberration in each region. For example, if the aberrant threshold is low (e.g. z-score of 2), then the amount threshold may be selected to be high (e.g., 15 matching regions or 80%). But, if the aberrant threshold is high (e.g., a z-score of 3), then the amount threshold may be lower (e.g., 5 matching regions or 60%). The amount of regions showing an aberration can also be a weighted value, e.g., one region that shows a high imbalance can be weighted higher than a region that just shows a little imbalance (i.e. there are more classifications than just positive and negative for the aberration). Such a weighting can act in a similar manner as certain regions that are required to have an aberration for the type of cancer to be identified.

[0102] In some embodiments, the threshold can be determined dynamically based on the number of matching regions for other types of cancers. For example, the threshold can be that the number of matching regions for the identified cancer be at least a specific number greater than the matching regions for the next most likely cancer type. Such a threshold can be an additional criterion in addition to a minimum threshold. Thus, in some instances, no cancer type might be identified if a sufficient number of matching regions do not exist.

B. Results

[0103] Method 600 was tested for a plurality of cancer types to determine the accuracy. Method 600 was tested with patients of known cancer type. Further, the thresholds used can be determined using samples of known cancer types. Different thresholds can be used for different cancer types.

[0104] The plasma DNA of each of 17 cancer patients (6 patients with HCC, 4 with colorectal cancers (CRC), 3 with breast cancers (BrC), 2 with lung cancers (LC) and 2 with nasopharyngeal carcinoma (NPC)) was sequenced. Copy number aberrations (CNAs) for each chromosome arm were analyzed for each patient based on the CAZA approach.

[0105] FIG. 7 shows chromosome arms that exhibit different patterns for different types of cancers in table 700 according to embodiments. CNAs that occur in $\geq 50\%$ of the cases are highlighted in color. Copy number losses are highlighted in red, and copy number gains are highlighted in green.

[0106] Table 700 has the chromosome arm listed on the leftmost column. Each of the other columns lists the type of cancer and a patient number for the cancer type. A deletion is

identified with '-'. An amplification is identified with '+'. A normal region is identified with 'Nil'.

[0107] The patterns of CNAs observed in the plasma samples are different for patients suffering from different types of cancers. Based on the most common patterns of CNAs observed in plasma, embodiments can deduce the potential tissue origin of cancers in patients with CNAs observed in plasma but the source of CNAs is unknown. The patterns of CNAs listed in table 700 are for illustration purposes, and a more comprehensive list of CNAs can be established by analyzing a much larger number of clinical samples.

[0108] The reference patterns of CNAs can also be determined from the analysis of tumor tissue. As examples, gains on 1q, 6p, 8q and 17q, and losses on 4q, 8p, 13q, 16q and 17p are commonly detected in HCC tumor tissues (Moinzadeh P et al. Br J Cancer 2005;92:935-941). Gains on 5p, 8q, 9p, 13q and 20q, and losses on 8p, 9p, 17p and 18q are commonly detected in CRC tumor tissues (Farzana et al. PLoS One 2012;2:231968 and Lips EH et al. J Pathol 2007;212:269-77). Gains on 5p, 7p, 7q, 8q 14q, 17q and 20q, and losses on 3p, 8p, 9p, 13q and 18q are commonly detected in non-small cell lung cancer tissues whereas gains on 3q, 5p, 14q and 19q, and losses on 3p, 4p, 4q, 5q, 10p, 10q, 13q, 15q, 17p and 22q are commonly detected in small cell lung cancer tissues (Zhao X et al. Cancer Res 2005;65:5561-70). Gains on 1q, 8q, 17q and 20q, and losses on 4p, 5q, 8p, 11q and 13q are common in breast cancer tissues (Andre F et al. Clin Cancer Res 2009;15:441-51). The patterns of CNAs described here are serve as illustrative examples and are not intended to be the only patterns that can be used in methods described herein.

[0109] Based on the CNA patterns in this example, assume that plasma DNA sequencing was performed for the patient BrC2 for the purpose of cancer screening. CNAs, including copy number gains for 1q, 3q, 8q, and 14q and copy number losses for 2p, 2q, 3p, 4p, 7q, 8p, 9p, 11p, 12p, 12q, 16q, and 17p, were observed. The CNAs in her plasma matched 13 typical CNAs for breast cancers. In contrast, her CNAs only matched 3, 6, 4, and 1 typical CNAs of HCC, CRC, LC, and NPC, respectively. Therefore, based on the CNA pattern of her plasma DNA, the most likely cancer that she has is deduced to be breast cancer. The selected threshold can be used to determine if the number of CNAs observed is compatible with the typical CNAs of certain cancer types. In this example, a threshold of 7, 8, 9, 10, 11, 12, or 13 can be used to classify the CNAs as compatible with breast cancer. A percentage of matching regions can also be used. For example, a percentage of regions that match the commonly

aberrant regions can be used. The commonly aberrant regions can be defined as regions that have a particular aberration in more than 50% of the reference samples.

[0110] In other embodiments, other statistical approaches, for example, but not limited to hierarchical clustering, can be used to deduce the most likely cancer type a patient is having.

5 For example, each reference sample can be assigned a multidimensional data point, where each dimension corresponds to a different region. In one implementation, each dimension can be assigned a -1 (for a deletion), 0 (normal), or a 1 (for an amplification). Higher numbers could be possible for different levels of amplifications. The samples for a particular cancer type will cluster together, and a new sample can be assigned to a cluster. The threshold could
10 correspond to the metric used to determine which cluster (if any) the new sample should be assigned, where the assigned cluster corresponds to the identified cancer type for the sample. For example, a cluster may have a centroid corresponding to regions of the reference patterns of the cluster shared by at least a predetermined number of reference patterns of the cluster. The cluster may include a boundary that defines which test patterns lie inside of the cluster.
15 The boundary can have various shapes beyond simply spherical. The boundary can be determined as part of the clustering analysis when determining which reference patterns belong to which cluster, where references patterns farthest away from the centroid but within the cluster can define the boundary. The threshold for determining whether a test pattern is part of a cluster can be considered the distance from the centroid to the boundary in the
20 direction from the centroid to the test pattern.

[0111] In yet another embodiment, the relative likelihood of having different types of cancer can be determined. The CNA pattern of a patient can be compared against the likelihood of a CNA for each type of cancer. For example, a patient has a 1q gain would be compared against the probability of the 1q gain for different types of cancers. For illustration
25 purposes, we assume that a 1q gain may occur in 70% of HCC patients, 20% of LC patients, and 1% of CRC patients. With these likelihoods, an odds ratio can be determined based on the relative percentage of patients with different cancer types having the CNA. For instance, based on the 1q gain, the patient may be considered 3.5 times more likely to have HCC than LC and 70 times more likely to have HCC than CRC. An odds ratio for HCC to LC to CRC
30 may be 70:20:1. One of skill would understand that this odds ratio could be expressed in several different, yet equivalent, forms. Odds ratios for different CNAs at chromosome arms other than 1q can be determined as well. An overall odds ratio may then be calculated with the likelihoods or odds ratios at the individual CNAs. In other words, given a CNA pattern

from a patient and likelihoods of different types of cancer having the given CNA pattern, the likelihoods of the different types of cancer can be compared to each other in an overall odds ratio. Although this example used likelihoods of CNAs at different chromosome arms, likelihoods of CNAs at different subchromosomal regions other than chromosome arms can be used. In some embodiments, if no CNA is found at a chromosome arm or other subchromosomal region in a patient, the pattern of no CNAs can be compared against the likelihood of not finding a CNA at the chromosome arm or subchromosomal region for different types of cancer. The pattern of regions without CNAs from a patient can then be used to determine the likelihood of different types of cancer. In addition, combining the analysis of regions with CNAs and regions without CNAs can be used to determine the likelihood or relative likelihood of a type of cancer at a potentially higher accuracy than if only one type of region is used.

[0112] In another example, assume that the patient NPC1 has the plasma DNA sequenced. CNAs, including copy number gains for 2q, 12q, and 22q and copy number losses for 6q and 18q were observed. The CNA pattern of this patient matched four of the typical CNAs for NPC. By comparison, this CNA pattern matched 0, 2, 0, and 0 typical CNAs for the patterns of HCC, CRC, BrC, and LC. In another embodiment, the lack of the typical CNA for a cancer type can also be counted. For example, none of the typical CNAs for NPC were absent in this patient. In contrast, 7, 16, 13, and 8 typical CNAs for HCC, CRC, BrC and LC were absent in this patient. Therefore, the CNA pattern of this patient is not suggestive of HCC, CRC, BrC, and LC.

[0113] FIG. 8A, 8B, and 8C show how the accuracy of this approach can further be enhanced by using higher resolution CNA analysis in table 800. The CNA affecting 1-Mb regions were identified in this cohort of cancer patients. Table 800 has the genomic coordinates of the 1-MB regions listed on the leftmost column. Each of the other columns lists the type of cancer and a patient number for the cancer type. A deletion is identified with '-'. An amplification is identified with '+'. A normal region is identified with 'Nil'.

[0114] In this example, the CNAs that spanned 1 Mb and were present in all the patients having the same cancer type were identified. With the higher resolution, subchromosomal CNAs that are present in a high proportion of patients with the same type of cancers can be identified. These cancer-type-specific CNAs are not identified in the arm-based analysis. For example, copy number gains on chromosome 18 spanning coordinates 30-31 Mb and 44-45

Mb were identified in all the three patients with lung cancer but were uncommon in patients with other cancer types. As discussed above, different statistical tests can be used to determine which cancer-specific CNA pattern is most similar to the tested case. Different statistical tests may include, for example, counting the number of typical CNAs in different cancer-associated CNA pattern and hierarchical clustering.

IV. SIZE ANALYSIS OF TUMOR-DERIVED DNA FRAGMENTS IN PLASMA

[0115] A statistically significant difference in the size distribution of DNA fragments can be used to identify an aberration, in a similar manner that the number of counts can. It has been reported that the size distribution of the total (i.e. tumoral plus non-tumoral) plasma DNA is increased in cancer patients (Wang BG, et al. Cancer Res. 2003; 63: 3966-8). However, if one is specifically studying the tumor-derived DNA (instead of the total (i.e. tumor plus non-tumor) amount of DNA), then it has been observed that the size distribution of tumor-derived DNA molecules is shorter than that of molecules derived from non-tumor cells (Diehl et al. Proc Natl Acad Sci U S A. 2005;102:16368-73). Therefore, the size distribution of circulating DNA can be used for determining if cancer-associated chromosomal aberrations are present.

[0116] The size analysis can use various parameters, as mentioned herein, and in U.S. Patent No. 8,620,593. For example, the Q or F values from above may be used. Such size values do not need a normalization by counts from other regions as these values do not scale with the number of reads. Techniques involving the depth and refinement of a region may be used. In some embodiments, a GC bias for a particular region can be taken into account when comparing two regions. In some implementations, the size analysis uses only DNA molecules.

A. Method

[0117] FIG. 9 is a flowchart illustrating a method 900 of analyzing a biological sample of an organism according to embodiments of the present invention. The biological sample may include nucleic acid molecules originating from normal cells and potentially from cells associated with cancer. At least some of the nucleic acid molecules may be cell free in the biological sample. In one aspect, method 900 can be directed to determining a classification of a sequence imbalance based on a separation value (e.g. a difference or ratio) for the size of fragments of a first chromosome and the size of fragments of one or more reference chromosomes.

[0118] In step 910, for each of a plurality of nucleic acid molecules in the biological sample, a size of the nucleic acid molecule may be measured. Obtaining the size of a nucleic acid molecule is described in U.S. Patent Publication No. 2013/0237431 entitled “Size-Based Analysis of Fetal DNA Fraction in Maternal Plasma” by Lo et al. filed March 7, 2013, the contents of which are incorporated herein by reference for all purposes.

[0119] In step 920, a location of the nucleic acid molecule in a reference genome of the organism may be identified. The location can be any part of a genome, as is described for step 120 and elsewhere. For example, it is identified which chromosome each of the plurality of nucleic acid molecules is derived. This determination can be made by a mapping to a reference genome.

[0120] In step 930, for each of the plurality of chromosomal regions, a respective group of nucleic acid molecules may be identified as being from a first chromosomal region based on the identified locations. The first chromosomal region may include a plurality of first loci.

[0121] In step 940, a computer system may calculate a first statistical value of a size distribution of the first group of nucleic acid molecules. In embodiments, the first statistical value may be determined by computing an area under a first curve at a specified size. The first curve may be a plot of a cumulative frequency of nucleic acid molecules for the first chromosomal region over a range of sizes. In one embodiment, the first statistical value can be an average, mean, median, or mode of the size distribution of the fragments corresponding to the first chromosome. In another embodiment, the first statistical value can include a sum of the length of fragments below a first size, which can be a type of cutoff. For example, each of the fragments that are smaller than 200 bp can have their lengths summed. The sum can be divided by another number, such as a sum of the lengths of all fragments corresponding to the first chromosome or a sum of the lengths of fragments greater than a second size cutoff (which may be the same as the first size). For example, the first statistical value can be a ratio of the total length of fragments below a first size cutoff relative to a total length of fragments, or a ratio of the total length of small fragments relative to a total length of large fragments.

[0122] In step 950, the first statistical value may be compared to a first reference value to determine a classification of whether the first chromosomal region exhibits an aberration. In embodiments, the first reference value may be a statistical value of a size distribution of a second group of nucleic acid molecules of a second chromosomal region. The second chromosomal region may be considered a reference chromosomal region. The first reference

value may be determined by computing an area under a second curve at the specified size. The second curve may be a plot of cumulative frequency of nucleic acid molecules for the second chromosomal region over the range of sizes. In one embodiment, the first reference value may be a statistical value for a plurality of reference chromosomes. In one

5 implementation, the statistical values can be combined such that the statistical value could be of one or more second chromosomes. In another embodiment, the statistical values for the plurality of reference chromosomes may be compared individually. The comparison may determine a classification of whether the first chromosomal region exhibits a deletion or an amplification.

10 **[0123]** The first statistical value and the first reference value may be compared to obtain a separation value. In one embodiment, the separation value can be a difference between the first statistical value and the first reference value is determined. In another embodiment, the separation value can be a ratio of the first statistical value to the first reference value. In yet another embodiment, a plurality of separation values can be determined, e.g., one for each
15 reference value, which can be calculated for each reference chromosome.

[0124] The separation value may be a difference in the proportion of short DNA fragments between the first chromosomal region and the reference chromosomal region using the following equation:

$$\Delta F = P(\leq 150bp)_{test} - P(\leq 150bp)_{ref}$$

20 where $P(\leq 150bp)_{test}$ denotes the proportion of sequenced fragments originating from the first chromosomal region with sizes ≤ 150 bp, and $P(\leq 150bp)_{ref}$ denotes the proportion of sequenced fragments originating from the reference chromosomal region with sizes ≤ 150 bp. In other embodiments, other size thresholds can be used, for example, but not limited to 100 bp, 110 bp, 120 bp, 130 bp, 140 bp, 160 bp and 166 bp. In other embodiments, the size
25 thresholds can be expressed in bases, or nucleotides, or other units. In some implementations, the reference chromosomal region can be defined as all the subchromosomal regions excluding the first chromosomal region. In other implementations, the reference region can be just a portion of the subchromosomal regions excluding the first chromosomal region.

[0125] The same groups of controls used in the count-based analysis can be used in the size-based analysis. A size-based z-score of the tested region can be calculated using the mean and SD values of ΔF of the controls:

$$\text{Size-based z-score} = \frac{\Delta F_{\text{sample}} - \text{mean } \Delta F_{\text{control}}}{SD \Delta F_{\text{control}}}.$$

- 5 **[0126]** The separation value may be compared to one or more cutoff values. In one embodiment, the comparison can be performed for each of a plurality of separation values. For example, a different separation value can be determined between the first statistical value and each reference value. In various implementations, each separation value can be compared to the same or different cutoff values. In another embodiment, a separation value is compared
- 10 to two cutoff values to determine whether the separation value is within a particular range. The range can include one cutoff to determine if a non-normal data point occurs (e.g. an aberration) and a second cutoff could be used to determine if the data point is likely caused by an error in measurement or analysis (e.g., if the separation value is larger than ever would be expected, even for a diseased sample).
- 15 **[0127]** A classification of whether a sequence imbalance (e.g. an aberration) exists for the first genomic location is determined based on the comparison. In one embodiment, a plurality of cutoffs (e.g. N cutoffs) can be used for a single separation value. In such an embodiment, N+1 classifications can be determined. For example, two cutoffs may be used to determine the classifications whether the chromosomal region is normal or healthy, indeterminate, or
- 20 aberrant (e.g. amplification or deletion). In another embodiment where a plurality of comparisons are performed (e.g. one for each separation value), the classification can be based on each of the comparisons. For example, a rule-based method can look at the classifications resulting from each of the comparisons. In one implementation, a definitive classification is only provided when all of the classifications are consistent. In another
- 25 implementation, the majority classification is used. In yet another implementation, a more complicated formula may be used based on how close each of the separation values is to a respective cutoff value, and these closeness values can be analyzed to determine a classification. For example, the closeness values could be summed (along with other factors, such as a normalization) and the result could be compared to another cutoff value. In other
- 30 embodiments, variations of method 900 can also be applied to a direct comparison of a

statistical value for the first chromosome to a cutoff value, which can be derived from a reference sample.

B. Correlation of size to cancer

[0128] For further analyses, we separately explored plasma DNA molecules of three different size groups, namely, those less than 150 bp, those between 150 and 180 bp, and those above 180 bp. There is a positive correlation (Pearson's $r = 0.6$; $p\text{-value} < 0.001$) between the proportion of DNA fragments less than 150 bp and the tumor DNA fraction in plasma (FIG. 10A). The tumor DNA fraction in FIGs. 10A, 10B, and 10C is shown in a logarithmic scale. No correlation ($r = -0.07$; $p\text{-value} = 0.95$) was observed between the proportion of DNA fragments with sizes between 150 and 180 bp and tumor DNA fraction in plasma (FIG. 10B). A negative correlation ($r = -0.41$; $p\text{-value} < -0.001$) was observed between the proportion of DNA more than 180 bp and tumor DNA fraction in plasma (FIG. 10C).

[0129] A lower tumor DNA fraction would more likely occur at the early stages of cancer, and a higher tumor DNA fraction would more likely occur at later stages of cancer. Thus, the existence of a larger average size (or other statistical value) than normal for DNA fragments can indicate an early-stage cancer, and existence of a smaller average size than normal for DNA fragments indicate a later stage cancer.

[0130] In other embodiments, the tumor DNA fraction can be measured. When the tumor DNA fraction is below a certain threshold, a size analysis can be performed to determine whether a statistical value of a size distribution is greater than a threshold (i.e., test whether the DNA fragments are long). When the tumor DNA fraction is above a certain threshold, a size analysis can be performed to determine whether a statistical value of a size distribution is less than a threshold (i.e., test whether the DNA fragments are short).

[0131] Methods of size analysis and data regarding the relationship of size with cancer are discussed in U.S. Patent Publication No. 2013/0040824 entitled "Detection of Genetic or Molecular Aberrations Associated with Cancer" by Lo et al. filed November 30, 2011, the contents of which are incorporated herein by reference for all purposes.

V. CONFIRMING CNA ABERRATION WITH SIZE ANALYSIS

[0132] We used massively parallel sequencing to study the size profiles of plasma DNA samples at single base resolution and in a genomewide manner. We used CAZA to identify tumor-derived plasma DNA for studying their specific size profiles.

5 [0133] In this study, we used the CAZA approach to identify chromosomal arms that showed plasma DNA quantitative aberrations suggestive of the presence of tumor-associated CNA. After identifying the chromosome arms with amplifications or deletions, we focused on these regions as a strategy to compare tumor-derived (enriched in the amplified regions) and non-tumor derived plasma DNA (enriched in the deleted regions). We believe that this
10 approach may provide a more robust means to identify tumoral DNA for size profiling analysis than based on the detection of cancer-associated mutations. For the latter, on average, it has been reported that there are of the order of thousands of point mutations in cancer genomes (29-32, 39). For CAZA, on the other hand, any of the myriad of plasma DNA molecules derived from the genomic regions exhibiting CNAs, totaling in terms of tens of
15 megabases, would be useful.

A. Combined Analysis

[0134] FIG. 11 shows a schematic illustration of the principle of plasma DNA size analysis in cancer patients. FIG. 11 shows stages 1110-2150. Stage 1110 shows the cells of the tissues in plasma. The tumor cells can include amplifications and/or deletions in various regions, as
20 is described above. The example shows one region amplified on a particular chromosome and another region deleted.

[0135] At stage 1120, the plasma is shown with contributions from various regions. DNA fragments are shown in the plasma sample. In cancer patients, plasma DNA is derived from both tumor (red molecules) and non-tumor cells (blue molecules). Genomic regions that are
25 amplified in the tumor tissue would contribute more tumoral DNA to plasma. Genomic regions that are deleted in the tumor tissue would contribute less DNA to plasma.

[0136] At stage 1130, paired-end sequencing is performed. The paired-end sequencing can be used to determine sizes of the DNA fragments in the plasma sample.

[0137] At stage 1140, a count-based analysis is used to identify aberrant regions. In the
30 example shown, a CAZA analysis was used to determine if a chromosome arm is over- or

under-represented in plasma DNA, suggestive of the presence of amplification or deletion of the chromosome arm in the tumor. A large positive z-score may indicate the presence of an amplification of the chromosome arm, while a large negative z-score may indicate the presence of a deletion of the chromosome arm. Other sizes of regions can be used besides the arms.

[0138] At stage 1150, the size distribution of a test region can be analyzed. As explained above, the tumor DNA fragments are shorter than DNA fragments of healthy cells. The DNA fragments of an aberrant region can be tested to confirm that the size analysis also shows a same aberration. In the example shown, a size distribution of a region exhibiting an amplification is compared to a size distribution of a region exhibiting a deletion. Thus, in some embodiments, the size profiles of plasma DNA molecules originating from chromosome arms that are under-represented (enriched for non-tumor DNA) and over-represented (enriched for tumor-derived DNA) can be compared, as described in greater detail below.

B. Size Difference between Two Regions

[0139] To compare the size profiles of plasma DNA originating from tumor and non-tumor tissues, we analyzed the plasma DNA fragments from the chromosome arms with CNAs. Based on previous studies (34-36) as well as our findings in this study, typical CNAs associated with HCC include 1p and 8p deletions, and 1q and 8q amplifications. A HCC case (H291) with 53% tumor-derived DNA in plasma is used to illustrate the principle. This case showed 8p deletion and 8q amplification in plasma. Thus, the tumor would release more plasma DNA from the amplified region of 8q than the deleted region of 8p. As a result, 8q would be relatively enriched for tumor-derived DNA and 8p would be relatively depleted of tumor DNA (or in other words, relatively enriched for non-tumor DNA) compared with regions without CNA. The size profiles of plasma DNA for 8p and 8q are shown in FIG. 12A. The size profile for 8q was on the left side of that for 8p, indicating that the size distribution of plasma DNA for 8q was shorter than that for 8p. Because 8q is enriched with tumor DNA, the data suggest that DNA released by the tumor tends to be shorter than DNA not originating from the tumor.

[0140] To quantify the degree of shortening, cumulative frequency plots (FIG. 12B) for the size profiles for 8p and 8q were constructed for each plasma sample. These plots show the progressive accumulation of DNA molecules, from short to long sizes, as a proportion of all

the plasma DNA molecules in the sample. The difference in the two curves ΔS (FIG. 12C) was then calculated as

$$\Delta S = S_{8q} - S_{8p}$$

where ΔS represents the difference in the cumulative frequencies between 8p and 8q at a particular size, and S_{8p} and S_{8q} represent the proportions of plasma DNA fragments less than a particular size on 8p and 8q, respectively. A positive value of ΔS for a particular size indicates a higher abundance of DNA shorter than that particular size on 8q compared with 8p. Using this method, we scanned the ΔS values from 50 bp to 250 bp for all HCC cases that exhibited CNAs on 8p and 8q in plasma. The difference in cumulative frequencies, ΔS , between 8q and 8p for the HCC case H291 is plotted as a red line in FIG. 12C. Compared with the healthy controls (grey lines), all these HCC cases showed higher abundance of plasma DNA shorter than 200 bp originating from 8q (enriched for tumor DNA) than from 8p (enriched for non-tumor DNA) (FIG. 13A). FIG. 13A shows a plot of ΔS against size for all the HCC cases with different CNAs on 8p and 8q in plasma. Cases with different ranges of fractional tumor DNA concentrations in plasma are shown in different colors. As the fractional tumor DNA concentration increases, the ΔS increases, indicating a higher abundance of shorter DNA fragments. These data further support that tumor-derived DNA was shorter than that of non-tumor derived DNA.

[0141] The value of ΔS attained a maximum at 166 bp suggesting that the key difference between plasma DNA derived from tumor and non-tumor tissues is the relative abundance of DNA < 166 bp and ≥ 166 bp. We denote this value as ΔS_{166} . The ΔS_{166} was plotted for all subjects of this study, including the HBV carriers and patients with liver cirrhosis (FIG. 13B). For the HCC group, patients with and without different CNAs on 8p and 8q as determined by plasma CAZA analysis are represented by red and black dots, respectively. For almost all of the non-HCC subjects, the ΔS_{166} values were close to 0 indicating that the size distributions for DNA from 8p and 8q were similar. The ΔS_{166} (or the value at some other specified size) can be compared to a threshold, and if the difference exceeds the threshold, then at least one of the regions can be identified as exhibiting an aberration. If one region is known to not have an aberration (e.g., from CNA analysis), then the other region would be identified as exhibiting an aberration when the difference exceeds a threshold. In such an embodiment, the sign of the difference can indicate the type of aberration. For example, when the first region has an amplification and the second region does not, then the difference would be a positive

number. When the first region has a deletion and the second region does not, then the difference would be a negative number. If an aberration is determined, then both regions can be identified as potentially having an aberration, with the sign indicating the type of aberration that each region may have. If the difference is big enough, it can indicate that one region has an amplification and the other region has a deletion (or amount of amplification is different), as then the difference would be larger than an amplified region compared to a normal region. The copy number analysis can provide an initial classification for the regions, so that a suitable threshold may be chosen.

[0142] Size analysis based on the plasma DNA size profiles of 1p and 1q was also performed (FIG. 14 and 15) and showed the same trend. In FIG. 15, for the HCC group, patients with and without different CNAs on 1p and 1q as determined by plasma CAZA analysis are represented by red and black dots, respectively. This size analysis can be performed using amplified region in a normal region, or normal region and a deleted region.

[0143] In another embodiment, a size distribution for amplified or deleted region can be compared to a size distribution of one or more reference subjects that are known to have cancer or known to be healthy. The size distribution can be represented by a value, e.g., a statistical value, such as a mean or median size.

[0144] Accordingly, the aberration of a chromosomal region can be used to select particular regions for a size analysis. The size analysis of the selected regions can then be used to determine a classification of a level of cancer. The combination of using CNA and size analysis can provide greater accuracy. The CNA analysis can occasionally yield false positives, i.e., patients who do not have cancer but who have regions with copy number aberration. Thus, a patient that is identified to have cancer due to a sufficient number of regions exhibiting aberration can then be confirmed using a size analysis. In one embodiment, the selected regions are ones that have amplification.

[0145] This study was designed with an intent to explore the plasma DNA size profile of HCC patients in a high resolution and comprehensive manner which may shed light on the mechanisms related to the generation or release of plasma DNA by tumor tissues. Another goal of the study was to resolve some of the apparent inconsistencies that existed in the literature regarding cancer-associated plasma DNA size profiles. Studies have reported the presence of longer DNA in the plasma of cancer patients (20-23) while others reported higher prevalence of cancer-associated DNA mutations among the shorter plasma DNA molecules

(12, 25). To achieve these study goals, a two-step approach was adopted. First, we measured the lengths of all DNA molecules in plasma samples of the recruited subjects with the use of paired-end massively parallel sequencing. This approach allows one to determine the lengths of individual plasma DNA molecules up to single base resolution. Furthermore, plasma DNA molecules across the genome could be analyzed and the relative amounts between DNA of different sizes could be determined with high precision. Hence, a broad and deep survey of the plasma DNA size profile could be obtained. Second, we took advantage of the relative difference in tumoral DNA content in plasma DNA originating from genomic locations that were associated with amplifications or deletions, the CAZA approach, as a means to identify tumor-derived plasma DNA for detailed analysis.

[0146] This study provides a number of insights into the biological mechanisms that might be involved in the release of plasma DNA. Plasma DNA of all recruited subjects, including the HBV carriers, patients with liver cirrhosis or HCC, exhibited a prominent peak at 166 bp (FIG. 14 and 16). This pattern is analogous to observations in the plasma of pregnant women and organ transplant recipients (26, 27). The presence of the characteristic 166 bp peak in the plasma DNA size profile of all groups of patients studied suggests that most of the circulating DNA molecules in human plasma, including that of pregnant women, transplant recipients, patients with HCC, liver cirrhosis or chronic HBV, resemble mononucleosomal units and are likely to originate from the process of apoptosis.

[0147] The study of the size profile of plasma DNA molecules bearing tumor-associated CNAs indicates that such molecules are shorter than those not carrying such signatures (FIG. 13). This is consistent with our observation that with increasing fractional concentrations of tumor DNA in plasma, the size profile of plasma DNA would shift towards the left. However, the fact that HCC patients with low fractional concentrations of tumor DNA in plasma had an apparently longer size distribution than healthy controls suggest that there was an additional component of plasma DNA that did not carry the tumor-associated genomic signatures. It is possible that this component would be derived from the non-neoplastic liver tissues surrounding the tumor. These long DNA molecules could be derived from necrosis instead of apoptosis. It has been reported that cell death associated with tissue necrosis may generate longer DNA fragments in addition to the typical oligonucleosomal DNA fragments (37, 38). For future studies, it would be interesting to study the DNA methylation profile of these longer DNA molecules to see if they bear resemblances to that expected for the liver.

[0148] We showed that populations of aberrantly short and long DNA molecules co-existed in the plasma of patients with hepatocellular carcinoma. The short ones preferentially carried the tumor-associated copy number aberrations.

[0149] In summary, we profiled the size distribution of plasma DNA in patients with HCC at single-nucleotide resolution. We have demonstrated a difference in the size of plasma DNA derived from tumor and non-tumor tissues.

[0150] The relationship between ΔS and tumor size was also analyzed. The plasma DNA samples of 10 HCC patients with 8p deletion and 8q amplification in plasma were analyzed using ΔS analysis. The ΔS was determined for the size difference between the plasma DNA fragments mapping to 8p and 8q. A positive value for ΔS indicates the more abundance of short DNA fragments below 150 bp for 8q compared with 8p. In FIG. 30, the values of ΔS were plotted against the longest dimension of the tumor of the HCC patients.

[0151] A positive correlation between ΔS and tumor size was observed ($r=0.876$, Pearson correlation). This observation suggests that the size distribution of plasma DNA fragments from regions exhibiting different types of CNAs can be used to reflect the size of the tumor in HCC patients.

[0152] The overall size distribution of the total plasma DNA was also analyzed for these 10 HCC patients. The percentage of plasma DNA fragments of less than 150 bp ($P(<150)$) was determined for each case and plotted against tumor size in FIG. 31. The proportion of short fragments was significantly higher in patients with larger cancer of more than 3 cm in the largest dimension. In one embodiment, the proportion of short fragments can be used to reflect the size and severity of the cancer. In other implementations, other cutoffs for size can be used, for example, but not limited to 100 bp, 110 bp, 120 bp, 130 bp, 140 bp, 160 bp and 166 bp.

[0153] A calibration function may be used to provide a relationship between size of the tumor and a statistical value. The calibration function may be determined from calibration data points of reference samples from organisms with tumors of known size. The calibration data point may include a measurement of the size of the tumor and a corresponding statistical measurement of sizes of nucleic acid molecules from a chromosomal region. When a new sample is obtained from a new subject, the statistical value may be determined, and the calibration function may be used to convert the statistical value into a tumor size. An example of a calibration function is a linear fit, similar to the linear fit shown in FIG. 30. Other types

of regression analysis, such as a least squares fit, may be used to generate the calibration function.

[0154] The calibration function be defined in a variety of ways, e.g., as a plurality of coefficients of a specified function, such as a linear or non-linear function. Other
5 embodiments can store a plurality of calibration data points (e.g., data points of the calibration function) so that the calibration function can be generated. Further, an interpolation can be performed between such calibration data points to obtain the calibration function. The calibration function may be stored in and retrieved from computer memory.

C. Method

10 **[0155]** FIG. 16 is a flowchart illustrating a method 1600 of performing CAZA and size analysis in order to analyze a biological sample of an organism according to embodiments of the present invention.

[0156] In step 1605, a plurality of chromosomal regions of an organism may be identified. Each chromosomal region may include a plurality of loci. One of the plurality of
15 chromosomal regions may be selected as a first chromosomal region. Identifying the plurality of chromosomal regions may be similar to step 610 of FIG. 6.

[0157] In step 1610, a location of a nucleic acid molecule in a reference genome of the organism may be identified for each of a plurality of nucleic acid molecules. Identifying the location of the nucleic acid molecule may be performed in a similar manner as step 620 of
20 FIG. 6.

[0158] In step 1615, a size of a nucleic acid molecule may be measured for each of the plurality of nucleic acid molecules in the biological sample. The size of the nucleic acid molecule may be measured similar to step 910 of FIG. 9.

[0159] In step 1620, a respective group of nucleic acid molecules may be identified, based
25 on the identified locations, as being from a chromosomal region for each chromosomal region of the plurality of chromosomal regions. The respective group may include at least one nucleic acid molecule located at each of the plurality of loci of the chromosomal region. Identification of the respective group of nucleic acid molecules may be similar to step 120 of FIG. 1.

[0160] In step 1625, a computer system may calculate a respective amount of the respective group of nucleic acid molecules. Calculating the respective amount may be similar to the calculation in step 130 of FIG. 1.

5 **[0161]** In step 1630, the respective amount may be compared to a count reference value to determine a count classification of whether the chromosomal region exhibits an amplification. Based on the comparison, the first chromosomal region may be identified as potentially exhibiting an aberration. Steps 1620-1630 may be performed in a similar manner as steps 120-140 of FIG. 1 or steps 630-650 of FIG. 6.

10 **[0162]** In step 1640, a first group of nucleic acid molecules may be identified as being from the first chromosomal region.

[0163] In step 1645, a computer system may calculate a first statistical value of a first size distribution of the first group of nucleic acid molecules. The first statistical value may be determined by computing an area under a first curve at a specified size. The first curve may be a plot of cumulative frequency of nucleic acid molecules for the first chromosomal region
15 over a range of sizes. Calculating the first statistical value in step 1645 may be similar to calculating the first statistical value in step 940 in FIG. 9.

[0164] In step 1650, the first statistical value may be compared to a size reference value to determine a size classification of whether the first chromosomal region exhibits an aberration. The size reference value may be determined by computing an area under a second curve at
20 the specified size. The second curve may be a plot of cumulative frequency of nucleic acid molecules for the second chromosomal region over the range of sizes. The comparison may be based on a difference between the two curves. In some embodiments, comparing the first statistical value to the size reference value may be similar to step 950 in FIG. 9.

[0165] In step 1655, a final classification of whether the first chromosomal region exhibits
25 an aberration may be determined. For example, at least one of the size classification and count classification can be used to determine whether the aberration exists for the first chromosomal region. In some embodiments, the final classification may be that the first aberration exists only when the count classification and the size classification indicate the same aberration. Thus, the comparison of the first statistical value to the size reference value
30 may confirm whether the first chromosomal region exhibits an aberration. In some embodiments, a set of size classifications may be determined for a set of chromosomal regions identified as aberrant based on corresponding count classifications. Based on the set

of size classifications, each of the chromosomal regions may be confirmed as aberrant or not aberrant.

[0166] In some embodiments, the final classification of whether the first chromosomal region exhibits an aberration may be based on multiple count reference values and multiple size reference values. Each of the count reference values can correspond to a different count classification (e.g., a discrimination between a unique pair of count classification, such as between level 1 and level 2, or between level 2 and level 3). Similarly, each of the size reference values can correspond to a different size classification. The final classification can be determined from the particular combination of size classification and count classification.

[0167] The size classification may include multiple classifications depending on a statistical value of the size distribution. For example, a large difference between the statistical value and a size reference value may result in a size classification corresponding to a high likelihood of an aberration, while a small difference between the statistical value and the size reference value may result in a size classification corresponding to a low likelihood of an aberration. Similarly, the count classification may include multiple classifications depending on the amount of a group of nucleic acid molecules. For example, a large difference between the amount of a group of nucleic acid molecules compared to a count reference value may result in a count classification corresponding to a high likelihood of an aberration, while a small difference may result in a count classification corresponding to a low likelihood of an aberration.

[0168] Accordingly, the final classification may be based on different thresholds for different size classifications and count classifications. For instance, a size classification indicating a high likelihood of an aberration may result in a final classification indicating an aberration given a count classification indicating a certain, possibly low, likelihood of an aberration. As the likelihood of an aberration as indicated by one of the size classification or the count classification increases, then the threshold for the likelihood indicated by the other classification is lowered. In some cases, one classification may show a high likelihood of a first type of aberration, the other classification may show a low likelihood of a second type of aberration, and the final classification may indicate that the first type of aberration is present. In some cases, the final classification may correspond to a likelihood or probability of an aberration.

D. Example Cases

[0169] The specificity of the detection of cancer-associated CNA can be improved by plasma DNA size analysis, as shown in the following two cases. Case 1 was a patient with hepatitis B-associated cirrhosis, and Case 2 was a chronic carrier of hepatitis B infection.

5 Both of them were not known of having any cancer at the time of recruitment. They had been followed clinically for two years since recruitment and no cancer was detected. Venous blood was collected from each of the two subjects at recruitment. The plasma DNA was sequenced. CNA involving chromosome 1q was detected in each of these two patients. For Case 1, the z-score for 1p and 1q were -2.3 and 15.5, respectively. These results are consistent with the
10 interpretation of 1q amplification. In the plasma DNA fragment size analysis, the ΔS was -0.019. The negative value of ΔS indicates that short DNA fragments were less abundant in 1q compared with 1p. As the count-based analysis suggests that 1q was amplified, the size-based analysis result is opposite to what we expected for cancer-associated CNAs. In cancer
15 patients, regions with copy number gain are expected to show an overall shorter size distribution due to the presence of more cancer-derived short fragments compared with regions with amplification or regions without any CNA. Therefore, the size analysis in this case is not suggestive of the presence of cancer-associated CNAs in the plasma DNA.

[0170] For Case 2, the z-scores for 1p and 1q were 0.4 and -4.4, respectively. These results are compatible with the interpretation of 1q deletion. In the plasma DNA fragment size
20 analysis, the ΔS was 0.044. The positive value of ΔS indicates that short DNA fragments were more abundant in 1q compared with 1p. As the count-based analysis suggests that 1q was deleted, the size-based analysis result is opposite to what we expected for cancer-associated CNAs. In cancer patients, regions with copy number loss are expected to show an overall longer size distribution due to the presence of less cancer-derived short fragments
25 compared with regions with amplification or regions without any CNA. Therefore, the size analysis in this case is not suggestive of the presence of cancer-associated CNAs in the plasma DNA.

VI. DETERMINATION OF STAGES OF CANCER

[0171] As mentioned above, the size of the DNA fragments can indicate the stage of the
30 cancer. A later stage of cancer exhibits smaller fragments for regions exhibiting amplification.

[0172] Apart from the intrinsic biological interest, plasma DNA size profiling may also be useful for the development of diagnostic approaches for detecting cancer-associated changes in plasma. For example, enrichment of tumoral DNA from plasma may be achieved by focusing on the analysis of short DNA fragments. In addition, we observed that the proportion of short DNA molecules bore a positive relationship with the fractional concentration of tumor-derived DNA in plasma. The changes in size profiles can be used for the monitoring of patients during the course of treatment. Furthermore, the presence of the population of long DNA molecules in the plasma of the patients with and without HCC warrants further investigation. When the tissue source or pathological process that governs the release of these DNA molecules are better understood, measuring the proportion of long DNA in plasma might be useful for the assessment of such diseases.

A. Plasma DNA size distribution of HCC patients

[0173] The size distributions of plasma DNA of the HCC patients, HBV carriers, cirrhosis patients and healthy controls are shown in FIG. 18 and 19. In FIG. 19, each individual is represented by a different color. In general, the most prominent peak was observed at 166 bp in the size distribution plot of each subject. This observation is consistent with previous reports on pregnant women and transplant recipients (26-28), suggesting that most of the circulating DNA molecules are derived from apoptosis. Interestingly, when compared with the median size distribution profile for 32 healthy controls (thick black line in FIG. 18), the sizes of plasma DNA in HCC patients with low fractional tumor DNA concentrations were longer. However, with increasing fractional concentrations of tumor DNA in plasma, the size distribution of plasma DNA shifted progressively to the left (FIG. 18).

[0174] As described earlier, FIG. 13A is a plot of ΔS against size for all the HCC cases with different CNAs on 8p and 8q in plasma. As the fractional tumor DNA concentration in plasma increases from less than 2% to over 8%, the ΔS increases, indicating a higher abundance of shorter DNA fragments. The fractional tumor DNA concentration in plasma may increase as the stage of cancer progresses. As a result, the amount of shorter DNA fragments may indicate a later stage of cancer. FIG. 13B shows that ΔS_{166} is higher for HCC patients, compared to non-HCC subjects, indicating that the relative abundance of DNA < 166 bp and ≥ 166 bp may be used to indicate the presence of cancer. Accordingly, ΔS_{166} may also indicate the stage of cancer.

[0175] FIG. 20 shows an example of when the proportion of short fragments can be used to differentiate HCC patients from healthy control subjects. The proportion of plasma DNA fragments less than 150 bp was plotted for 32 healthy subjects, HCC patients with tumor DNA fraction of less than 2% in plasma and HCC patients with tumor DNA fraction of greater than 6% in plasma. Compared with healthy control subjects (labeled as 'CTR'), HCC patients with tumor DNA fraction of less than 2% had significantly lower proportion of short DNA fragments of less than 150 bp ($p = 0.0002$, t-test), and those with tumor DNA fraction of greater than 6% had significantly higher proportion of short fragments ($p = 0.003$, t-test). HCC patients with a tumor DNA fraction from 2% to 6% have a proportion of DNA fragments between HCC patients with a tumor fraction of less than 2% and HCC patients with a tumor fraction greater than 6%. In this manner, HCC patients with the tumor fraction from 2% to 6% may have a distribution similar to the healthy control subjects.

[0176] FIG. 21 shows a receiver operating characteristic (ROC) curve for applying $P(<150)$ to differentiate HCC patients with less than 2% tumor DNA fraction from healthy control subjects. The tumor fraction was determined based on the magnitude of under-representation of the chromosome regions exhibiting under-representation in the plasma that were compatible with a copy number loss in the tumor. For cases without significant under-representation of any chromosome arm, the magnitude of over-representation for regions that were compatible with copy number gain was used to determine the tumor fraction with an assumption of single copy gain. The tumor fraction can be determined with the following equation:

$$\text{Tumor fraction} = \frac{|P_{test} - P_{normal}|}{P_{normal} \times \Delta N / 2}$$

where P_{test} represents the proportion of fragments mapped to the chromosome arm of interest for the test case, P_{normal} represents the mean proportion of fragments mapped to the chromosome arm for the healthy controls, and ΔN represents the magnitude of the copy number change (e.g, 1 for either a duplication or a deletion, and higher numbers for higher order amplifications). The area under the curve (AUC) was 0.776 with 95% confidence limits of 0.670 and 0.882. This result indicates that size analysis can be used to identify HCC patients with tumor fraction of less than 2% in plasma. ROC curve analysis indicates that different thresholds can be selected to achieve different sensitivities and specificities.

[0177] FIG. 22, similar to FIG. 21, shows that size analysis with $P(<150)$ can also detect HCC patients with a tumor fraction of greater than 6% in the plasma. The AUC for differentiating these patients from healthy subjects was 0.893 with 95% confidence limits of 0.761 and 1.000.

5 [0178] FIG. 23 shows that the proportion of long plasma DNA fragments can be used for detecting HCC, as FIG. 20 showed with the proportion short plasma DNA fragments. In this example, the proportion of fragments greater than 180 bp, denoted as $P(>180)$, was plotted for HCC patients with less than 2% and greater than 6% tumor DNA fraction in plasma and healthy control subjects. This proportion was significantly higher in HCC patients with less
10 than 2% tumor DNA fraction ($p < 0.00001$, t-test).

[0179] FIG. 24 shows an ROC curve for using $P(>180)$ to differentiate HCC patients with less than 2% tumor DNA fraction from healthy control subjects. The AUC was 0.883 with 95% confidence limits of 0.805 and 0.961.

[0180] FIG. 25 provides another example of the different size distributions of DNA
15 fragments with different tumor DNA fractions. FIG. 25 shows boxplots of the median fragment size of healthy control subjects, HCC patients with less than 2% tumor DNA fraction, and HCC patients with greater than 6% tumor DNA fraction. The median size of DNA fragments of the HCC patients with less than 2% tumor DNA fraction were significantly longer ($P < 0.00001$, t-test) than the healthy control subjects. In contrast, the
20 median size of DNA fragments of the HCC patients with greater than 6% tumor DNA fraction were significantly shorter ($p = 0.03$, t-test). FIG. 25 supports the use of DNA fragment size as a way to determine stage of cancer. A longer median size is associated with a smaller tumor DNA fraction, while a shorter median size is associated with a larger tumor DNA fraction. If an individual has a smaller tumor DNA fraction below a first cutoff and a
25 median size above a long size threshold, then early stage cancer may be confirmed. On the other hand, if an individual has a larger tumor DNA fraction above a second cutoff and a median size below a short size threshold, then late stage cancer may be confirmed.

[0181] HCC patients with a tumor DNA fraction from 2% to 6% have a median DNA fragments size between HCC patients with a tumor fraction of less than 2% and HCC patients
30 with a tumor fraction greater than 6%. In this manner, HCC patients with the tumor fraction from 2% to 6% may have a distribution similar to the healthy control subjects in FIG. 25. Hence, if an individual has a tumor DNA fraction from the low cutoff to the high cutoff and a

median size from a short size threshold to a long size threshold, then middle stage cancer may be confirmed.

[0182] FIGs. 26 and 27 are ROC curves that show that different size thresholds can be used to differentiate HCC patients from healthy control subjects. FIG. 26 is an ROC curve for using median fragment size to differentiate between HCC patients with less than 2% tumor DNA fraction and healthy control subjects. The AUC was 0.812 with 95% confidence limits of 0.718 and 0.907.

[0183] FIG. 27 is an ROC curve for using median fragment size to differentiate between HCC patients with greater than 2% tumor DNA fraction and healthy control subjects. The AUC was 0.795 with 95% confidence limits of 0.627 and 0.963.

[0184] Other statistical characteristics of the size distribution (e.g., median, mean, percentile) can be used as a parameter for the differentiation of HCC patients and healthy subjects.

[0185] In addition to analyzing the size distribution of plasma DNA fragments arising from all genomic regions, size analysis can also focus on DNA fragments arising from specific genomic regions. A specific genomic region may be a chromosome arm.

[0186] FIG. 28 shows a boxplot of the proportion of short plasma DNA fragments of less than 150 bp that were aligned to chromosome 1q for HCC patients with greater than 6% tumor DNA fraction and for healthy control subjects. The proportion of short fragments was significantly higher ($p < 0.00001$, t-test) in the HCC patients.

[0187] FIG. 29 is an ROC curve for using the proportion of short plasma DNA fragments of less than 150 bp to differentiate between HCC patients with greater than 6% tumor DNA fraction and healthy control subjects. The AUC was 0.915 with a 95% confidence interval from 0.808 to 1.000.

B. Method

[0188] FIG. 17 is a flowchart illustrating a method 1700 of analyzing a biological sample of an organism according to embodiments of the present invention. The biological sample may include nucleic acid molecules originating from normal cells and from cells associated with cancer. At least some of the nucleic acid molecules are cell-free in the biological sample.

[0189] In step 1710, for each of a plurality of the nucleic acid molecules in the biological sample, a size of the nucleic acid molecule is measured. The size of the nucleic acid molecule may be measured similar to step 910 of FIG. 9.

[0190] In step 1720, a location of the nucleic acid molecule in a reference genome of the organism is identified. Identifying the location of the nucleic acid molecule may be performed in a similar manner as step 620 of FIG. 6.

[0191] In step 1730, a first group of nucleic acid molecules is identified as being from a first chromosomal region based on the identified locations. The first chromosomal region may include a plurality of first loci. Identification of the respective group of nucleic acid molecules may be similar to step 120 of FIG. 1.

[0192] In step 1740, a computer system may calculate a first statistical value of a size distribution of the first group of nucleic acid molecules. Calculating the respective amount may be similar to the calculation in step 130 of FIG. 1.

[0193] In step 1750, a fraction of nucleic acid molecules originating from cells associated with cancer may be measured. The fraction may be calculated according to methods described in U.S. Patent Publication No. 2013/0040824 entitled "Detection of Genetic or Molecular Aberrations Associated with Cancer" by Lo et al. filed November 30, 2011. The fraction of tumor nucleic acid molecules corresponds to a proportion of the nucleic acid molecules in the sample that are from the tumor(s). The fraction/proportion may be expressed as any percentage or decimal value.

[0194] The following examples are methods for the measurement of the fraction of tumor nucleic acids but other methods can be used. The fraction of tumor nucleic acids can be determined based on the magnitude of under-representation (or over-representation) in the plasma for regions exhibiting significant under-representation that is compatible with copy number loss (or copy number gain) in the tumor tissues. Another example is to determine the degree of allelic imbalance on two homologous chromosomes for regions affected by copy number aberrations, e.g., regions with the loss of one copy of the two homologous chromosomes. Another example is to determine the fractional concentration of a cancer-associated mutation, including single nucleotide mutation, deletion of nucleotide(s), and translocation. The tumor fraction may be determined by methods described with FIG. 21 above.

[0195] In step 1760, a first reference value based on the measured fraction may be selected. In one example, selecting the first reference value may include selecting a size threshold when the measured fraction is below a cutoff. In another example, selecting the first reference value may include selecting a size threshold when the measured fraction is above a cutoff. In these examples, the cutoffs and the size thresholds may differ and may depend on the value of the measured fraction.

[0196] In step 1770, the first statistical value may be compared to a first reference value to determine a stage of cancer of the biological sample. The first statistical value may be any statistical value described herein.

[0197] Whether cancer exists can be confirmed based on the size analysis along with the measured fraction of nucleic acid molecules originating from cells associated with cancer. For example, when the measured fraction is below a low cutoff, it can be confirmed whether the size distribution is longer than for healthy controls (e.g., whether the first statistical value is above the size threshold). If the size distribution is longer than for healthy controls, this can confirm an early stage of cancer. Examples of the low cutoff are 0.01, 0.015, 0.02, or 0.025. As another example, when the measured fraction is above a high cutoff, it can be confirmed whether the size distribution is shorter than for healthy controls (e.g., whether the first statistical value is below the size threshold). If the size distribution is shorter for healthy controls, this can confirm a late stage of cancer. Examples of the high cutoff may be a fraction of 0.03, 0.035, 0.04, 0.045, 0.05, 0.055, 0.06, 0.065, or 0.07.

[0198] We showed that there were additional populations of shorter and longer DNA molecules in plasma of HCC patients. These data might have resolved the apparent inconsistencies that existed in the literature where groups reported the presence of either an increase in the longer or the shorter DNA molecules in the plasma of cancer patients.

VII. MATERIALS AND METHODS

[0199] Techniques used in obtaining the results of FIGs. 2-5 are now discussed. Such techniques can be used in other examples above.

[0200] Subjects recruited for study included 90 patients with HCC admitted to the Department of Surgery of the Prince of Wales Hospital, Hong Kong, for tumor resection. All blood samples were collected before operation. Sixty-seven HBV carriers and 36 patients with HBV-related cirrhosis were recruited from the Department of Medicine and

Therapeutics of the Prince of Wales Hospital, Hong Kong. All patients gave written informed consent and the study was approved by the institutional review board.

[0201] In order to extract DNA and prepare sequence libraries, peripheral blood samples were collected into EDTA-containing tubes. Peripheral blood samples were centrifuged at 1,600 g for 10 min at 4 °C. The plasma portion was recentrifuged at 16,000g for 10 min at 4 °C to obtain cell-free plasma. DNA was extracted from 3 to 4.8 mL of plasma using the QIAamp DSP DNA Blood Mini Kit (Qiagen). The plasma DNA was concentrated with a SpeedVac Concentrator (Savant DNA120; Thermo Scientific) into a 75-μL final volume per sample. Indexed DNA libraries were prepared by using the Kapa Library Preparation Kit (Kapa Biosystems) following the manufacturer's instructions. The adaptor-ligated DNA was enriched by a 14-cycle PCR using the KAPA HiFi HotStart ReadyMix PCR Kit (Kapa Biosystems). The libraries were then analyzed by a 2100 Bioanalyzer (Agilent) and quantified by the Kapa Library Quantification Kit (Kapa Biosystems) before sequencing.

[0202] To sequence and align DNA, each DNA library was diluted and hybridized to a paired-end sequencing flow cell (Illumina). DNA clusters were generated on a cBot cluster generation system (Illumina) with the TruSeq PE Cluster Generation Kit v3 (Illumina), followed by 76 × 2 cycles of sequencing on a HiSeq 2000 system (Illumina) with the TruSeq SBS Kit v3 (Illumina). Sequencing was performed using a 4-plex protocol. We performed an additional 7 cycles of sequencing to decode the index sequence on each sequenced DNA molecule. Real-time image analysis and base calling were performed using the HiSeq Control Software (HCS) v1.4 and Real Time Analysis (RTA) Software v1.13 (Illumina), by which the automated matrix and phasing calculations were based on the spiked-in PhiX control v3 sequenced with the libraries. After base calling, adapter sequences and low quality bases (i.e. quality score < 5) were removed.

[0203] For sequencing data analysis, sequences from each lane were assigned to the corresponding samples based on the six-base index sequences. The sequenced reads were then aligned to the non-repeat-masked human reference genome (NCBI build 37/hg19) using the Short Oligonucleotide Alignment Program 2 (SOAP2) (40). Up to two nucleotide mismatches were allowed for each member of the paired-end reads but insertions or deletions were not allowed. Reads mapped to a unique genomic location were used for downstream analyses. Paired-end reads aligned to the same chromosome with a correct orientation and spanning an insert size of ≤ 600 bp were retained for downstream size analyses. After

alignment to the reference human genome, the size of each plasma DNA fragment could be deduced from the coordinates of the nucleotides at the outermost ends of each pair of sequence reads. The first single-end reads were used for CNA analysis. Reads with mapping quality of greater than 30 (i.e. 1 erroneous alignment per 1,000 alignments) using the Bowtie 2 software (41) were accepted.

[0204] For performing CAZA analysis for CNA, the entire human genome was divided into 100-kb bins. The GC-corrected read count was determined for each 100-kb bin as reported previously (42). The number of GC-corrected read counts for each chromosome arm of interest was determined by summing all values of each 100-kb bin on the chromosome arm. A z-score statistic was used to determine if the plasma DNA representation in a chromosome arm would be significantly increased or decreased when compared with the reference group. The percentage of sequencing reads mapped to each chromosome arm was calculated and compared with the mean value of the 32 healthy control subjects for the respective chromosome arm. An arm-level z-score was calculated as

$$z - score = \frac{P_{test} - P_{normal}}{SD_{normal}}$$

where P_{test} represents the proportion of fragments mapped to the chromosome arm of interest for the test case; P_{normal} and SD_{normal} represent the mean and SD of the proportion of fragments mapped to the chromosome arm for the healthy controls, respectively. Chromosome arms with z scores of < -3 and > 3 were regarded as having CNAs in plasma corresponding to deletions and amplifications, respectively.

[0205] The fractional concentration of tumor-derived DNA in the plasma (F) can be calculated as

$$F = \frac{|P_{test} - P_{normal}|}{\Delta N / 2 \times P_{normal}}$$

where P_{test} represents the proportion of fragments mapped to the chromosome arm of interest for the test case; P_{normal} represents the mean proportion of fragments mapped to the chromosome arm for the healthy controls and ΔN represents the copy number change. For cases showing a deletion in at least one chromosome arm, we calculate F based on the deleted chromosome arm(s). As most chromosome arm deletions involve only one of the two homologous chromosomes (33), we assumed a single copy loss for our analysis. For the 24

cases with only chromosome arm amplification but no deletion, F was calculated based on the amplified arm with the assumption of single copy gain.

[0206] Sequencing data analysis was performed by using bioinformatics programs written in Perl and R languages. A p-value of < 0.05 was considered as statistically significant and all probabilities were two-tailed.

VIII. COMPUTER SYSTEM

[0207] Any of the computer systems mentioned herein may utilize any suitable number of subsystems. Examples of such subsystems are shown in FIG. 32 in computer apparatus 10. In some embodiments, a computer system includes a single computer apparatus, where the subsystems can be the components of the computer apparatus. In other embodiments, a computer system can include multiple computer apparatuses, each being a subsystem, with internal components. A computer system can include desktop and laptop computers, tablets, mobile phones and other mobile devices.

[0208] The subsystems shown in FIG. 32 are interconnected via a system bus 75.

Additional subsystems such as a printer 74, keyboard 78, storage device(s) 79, monitor 76, which is coupled to display adapter 82, and others are shown. Peripherals and input/output (I/O) devices, which couple to I/O controller 71, can be connected to the computer system by any number of means known in the art such as input/output (I/O) port 77 (e.g., USB, FireWire[®]). For example, I/O port 77 or external interface 81 (e.g. Ethernet, Wi-Fi, etc.) can be used to connect computer apparatus 10 to a wide area network such as the Internet, a mouse input device, or a scanner. The interconnection via system bus 75 allows the central processor 73 to communicate with each subsystem and to control the execution of instructions from system memory 72 or the storage device(s) 79 (e.g., a fixed disk, such as a hard drive or optical disk), as well as the exchange of information between subsystems. The system memory 72 and/or the storage device(s) 79 may embody a computer readable medium. Another subsystem is a data collection device 85, such as a camera, microphone, accelerometer, and the like. Any of the data mentioned herein can be output from one component to another component and can be output to the user.

[0209] A computer system can include a plurality of the same components or subsystems, e.g., connected together by external interface 81 or by an internal interface. In some embodiments, computer systems, subsystem, or apparatuses can communicate over a network. In such instances, one computer can be considered a client and another computer a server,

where each can be part of a same computer system. A client and a server can each include multiple systems, subsystems, or components.

[0210] It should be understood that any of the embodiments of the present invention can be implemented in the form of control logic using hardware (e.g. an application specific
5 integrated circuit or field programmable gate array) and/or using computer software with a generally programmable processor in a modular or integrated manner. As used herein, a processor includes a single-core processor, multi-core processor on a same integrated chip, or multiple processing units on a single circuit board or networked. Based on the disclosure and teachings provided herein, a person of ordinary skill in the art will know and appreciate other
10 ways and/or methods to implement embodiments of the present invention using hardware and a combination of hardware and software.

[0211] Any of the software components or functions described in this application may be implemented as software code to be executed by a processor using any suitable computer language such as, for example, Java, C, C++, C#, Objective-C, Swift, or scripting language
15 such as Perl or Python using, for example, conventional or object-oriented techniques. The software code may be stored as a series of instructions or commands on a computer readable medium for storage and/or transmission, suitable media include random access memory (RAM), a read only memory (ROM), a magnetic medium such as a hard-drive or a floppy disk, or an optical medium such as a compact disk (CD) or DVD (digital versatile disk), flash
20 memory, and the like. The computer readable medium may be any combination of such storage or transmission devices.

[0212] Such programs may also be encoded and transmitted using carrier signals adapted for transmission via wired, optical, and/or wireless networks conforming to a variety of protocols, including the Internet. As such, a computer readable medium according to an
25 embodiment of the present invention may be created using a data signal encoded with such programs. Computer readable media encoded with the program code may be packaged with a compatible device or provided separately from other devices (e.g., via Internet download). Any such computer readable medium may reside on or within a single computer product (e.g. a hard drive, a CD, or an entire computer system), and may be present on or within different
30 computer products within a system or network. A computer system may include a monitor, printer, or other suitable display for providing any of the results mentioned herein to a user.

[0213] Any of the methods described herein may be totally or partially performed with a computer system including one or more processors, which can be configured to perform the steps. Thus, embodiments can be directed to computer systems configured to perform the steps of any of the methods described herein, potentially with different components

5 performing a respective steps or a respective group of steps. Although presented as numbered steps, steps of methods herein can be performed at a same time or in a different order. Additionally, portions of these steps may be used with portions of other steps from other methods. Also, all or portions of a step may be optional. Additionally, any of the steps of any of the methods can be performed with modules, circuits, or other means for performing these
10 steps.

[0214] The specific details of particular embodiments may be combined in any suitable manner without departing from the spirit and scope of embodiments of the invention. However, other embodiments of the invention may be directed to specific embodiments relating to each individual aspect, or specific combinations of these individual aspects.

15 **[0215]** The above description of example embodiments of the invention has been presented for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form described, and many modifications and variations are possible in light of the teaching above.

[0216] A recitation of “a”, “an” or “the” is intended to mean “one or more” unless
20 specifically indicated to the contrary. The use of “or” is intended to mean an “inclusive or,” and not an “exclusive or” unless specifically indicated to the contrary.

[0217] All patents, patent applications, publications, and descriptions mentioned herein are incorporated by reference in their entirety for all purposes. None is admitted to be prior art.

IX. REFERENCES

1. Chan KCA (2013) Scanning for cancer genomic changes in plasma: toward an era of personalized blood-based tumor markers. *Clin Chem* 59(11):1553-1555.
2. Dawson SJ, Rosenfeld N, & Caldas C (2013) Circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med* 369(1):93-94.
3. Bidard FC, Weigelt B, & Reis-Filho JS (2013) Going with the flow: from circulating tumor cells to DNA. *Sci Transl Med* 5(207):207ps214.
4. Chan KCA, *et al.* (2013) Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. *Clin Chem* 59(1):211-224.
5. Heitzer E, *et al.* (2013) Establishment of tumor-specific copy number alterations from plasma DNA of patients with cancer. *Int J Cancer* 133(2):346-356.
6. Heitzer E, *et al.* (2013) Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. *Genome Med* 5(4):30.
7. Leary RJ, *et al.* (2012) Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci Transl Med* 4(162):162ra154.
8. Chan KCA, *et al.* (2013) Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proc Natl Acad Sci U S A* 110(47):18761-18768.
9. Chan KCA, *et al.* (2008) Quantitative analysis of circulating methylated DNA as a biomarker for hepatocellular carcinoma. *Clin Chem* 54(9):1528-1536.
10. Wong IH, *et al.* (1999) Detection of aberrant p16 methylation in the plasma and serum of liver cancer patients. *Cancer Res* 59(1):71-73.
11. Balgkouranidou I, *et al.* (2014) Breast cancer metastasis suppressor-1 promoter methylation in cell-free DNA provides prognostic information in non-small cell lung cancer. *Br J Cancer* 110(8):2054-2062.
12. Diehl F, *et al.* (2005) Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc Natl Acad Sci U S A* 102(45):16368-16373.

13. Yung TKF, *et al.* (2009) Single-molecule detection of epidermal growth factor receptor mutations in plasma by microfluidics digital PCR in non-small cell lung cancer patients. *Clin Cancer Res* 15(6):2076-2084.
14. Murtaza M, *et al.* (2013) Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* 497(7447):108-112.
15. Forshaw T, *et al.* (2012) Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci Transl Med* 4(136):136ra168.
16. Lo YMD, *et al.* (1999) Quantitative analysis of cell-free Epstein-Barr virus DNA in plasma of patients with nasopharyngeal carcinoma. *Cancer Res* 59(6):1188-1191.
17. Chan KCA, *et al.* (2013) Early detection of nasopharyngeal carcinoma by plasma Epstein-Barr virus DNA analysis in a surveillance program. *Cancer* 119(10):1838-1844.
18. McBride DJ, *et al.* (2010) Use of cancer-specific genomic rearrangements to quantify disease burden in plasma from patients with solid tumors. *Genes, Chromosomes & Cancer* 49(11):1062-1069.
19. Leary RJ, *et al.* (2010) Development of personalized tumor biomarkers using massively parallel sequencing. *Sci Transl Med* 2(20):20ra14.
20. Chan KCA, Leung SF, Yeung SW, Chan ATC, & Lo YMD (2008) Persistent aberrations in circulating DNA integrity after radiotherapy are associated with poor prognosis in nasopharyngeal carcinoma patients. *Clin Cancer Res* 14(13):4141-4145.
21. Gao YJ, *et al.* (2010) Increased integrity of circulating cell-free DNA in plasma of patients with acute leukemia. *Clin Chem Lab Med* 48(11):1651-1656.
22. Umetani N, *et al.* (2006) Increased integrity of free circulating DNA in sera of patients with colorectal or periampullary cancer: direct quantitative PCR for ALU repeats. *Clin Chem* 52(6):1062-1069.
23. Wang BG, *et al.* (2003) Increased plasma DNA integrity in cancer patients. *Cancer Res* 63(14):3966-3968.
24. Umetani N, *et al.* (2006) Prediction of breast tumor progression by integrity of free circulating DNA in serum. *J Clin Oncol* 24(26):4270-4276.

25. Schwarzenbach H, *et al.* (2012) Loss of heterozygosity at tumor suppressor genes detectable on fractionated circulating cell-free tumor DNA as indicator of breast cancer progression. *Clin Cancer Res* 18(20):5719-5730.
26. Lo YMD, *et al.* (2010) Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci Transl Med* 2(61):61ra91.
27. Zheng YWL, *et al.* (2012) Nonhematopoietically derived DNA is shorter than hematopoietically derived DNA in plasma: a transplantation model. *Clin Chem* 58(3):549-558.
28. Yu SCY, *et al.* (2014) Size-based molecular diagnostics using plasma DNA for noninvasive prenatal testing. *Proc Natl Acad Sci U S A* 111(23):8583-8588.
29. Pleasance ED, *et al.* (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 463(7278):191-196.
30. Fujimoto A, *et al.* (2012) Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat Genet* 44(7):760-764.
31. Tao Y, *et al.* (2011) Rapid growth of a hepatocellular carcinoma and the driving mutations revealed by cell-population genetic analysis of whole-genome data. *Proc Natl Acad Sci U S A* 108(29):12042-12047.
32. Totoki Y, *et al.* (2011) High-resolution characterization of a hepatocellular carcinoma genome. *Nat Genet* 43(5):464-469.
33. Beroukhim R, *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature* 463(7283):899-905.
34. Chiang DY, *et al.* (2008) Focal gains of VEGFA and molecular classification of hepatocellular carcinoma. *Cancer Res* 68(16):6779-6788.
35. Kan Z, *et al.* (2013) Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. *Genome Res* 23(9):1422-1433.
36. Kim TM, *et al.* (2008) Clinical implication of recurrent copy number alterations in hepatocellular carcinoma and putative oncogenes in recurrent gains on 1q. *Int J Cancer* 123(12):2808-2815.

37. Nakano H & Shinohara K (1994) X-ray-induced cell death: apoptosis and necrosis. *Radiation Research* 140(1):1-9.
38. Walker NI, Harmon BV, Gobe GC, & Kerr JF (1988) Patterns of cell death. *Methods and Achievements in Experimental Pathology* 13:18-54.
39. Alexandrov LB, *et al.* (2013) Signatures of mutational processes in human cancer. *Nature* 500(7463):415-421.
40. Li R, *et al.* (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25(15):1966-1967.
41. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9(4):357-359.
42. Chen EZ, *et al.* (2011) Noninvasive prenatal diagnosis of fetal trisomy 18 and trisomy 13 by maternal plasma DNA sequencing. *PLoS One* 6(7):e21791.

WHAT IS CLAIMED IS:

1. A method of analyzing a biological sample of an organism, the biological sample including nucleic acid molecules originating from normal cells and potentially from cells associated with cancer, wherein at least some of the nucleic acid molecules are cell-free in the biological sample, the method comprising:
 - identifying a plurality of chromosomal regions of the organism, each chromosomal region including a plurality of loci;
 - for each of a plurality of the nucleic acid molecules in the biological sample:
 - measuring a size of the nucleic acid molecule; and
 - identifying a location of the nucleic acid molecule in a reference genome of the organism;
 - for each of the plurality of chromosomal regions:
 - identifying a respective group of nucleic acid molecules as being from the chromosomal region based on the identified locations, the respective group including at least one nucleic acid molecule located at each of the plurality of loci of the chromosomal region;
 - calculating, with a computer system, a respective amount of the respective group of nucleic acid molecules; and
 - comparing the respective amount to a count reference value to determine a count classification of whether the chromosomal region exhibits an aberration, the count classification for a first chromosomal region of the plurality of chromosomal regions indicating a first aberration;
 - identifying a first group of nucleic acid molecules as being from the first chromosomal region based on the identified locations;
 - calculating, with a computer system, a first statistical value of a first size distribution of the first group of nucleic acid molecules; and
 - comparing the first statistical value to a size reference value to determine a size classification of whether the first chromosomal region exhibits the first aberration; and
 - determining a final classification of whether the first chromosomal region exhibits the first aberration using the count classification and the size classification of the first chromosomal region.

2. The method of claim 1, wherein the final classification is that the first aberration exists only when the count classification and the size classification indicate a same aberration.

3. The method of claim 1, wherein the plurality of chromosomal regions are non-overlapping.

4. The method of claim 1, wherein the final classification is that the first aberration exists, the method further comprising:

obtaining a calibration function that provides a relationship between a size of a tumor and the first statistical value; and

determining the size of the tumor using the calibration function.

5. The method of claim 4, wherein the calibration function is determined from calibration data points of reference samples from organisms with tumors of known size, wherein a calibration data point includes a measurement of the size of the tumor and a corresponding statistical measurement of sizes of nucleic acid molecules from the first chromosomal region.

6. The method of claim 1, wherein the respective value comprises a mean of the first size distribution, a median of the first size distribution, a mode of the of the first size distribution, or a proportion of nucleic acid molecules having a size below a size threshold.

7. The method of claim 1, wherein the size reference value corresponds to a second statistical value of a second size distribution of a second chromosomal region.

8. The method of claim 7, wherein comparing the first statistical value to the size reference value includes:

determining a separation value between the first statistical value and the size reference value; and

comparing the separation value to a threshold value.

9. The method of claim 1, further comprising:

determining a set of size classifications for a set of chromosomal regions identified as aberrant based on corresponding count classifications; and

confirming whether the set of chromosomal regions are aberrant based on the set of size classifications.

10. A method of analyzing a biological sample of an organism, the biological sample including nucleic acid molecules originating from normal cells and potentially from cells associated with cancer, wherein at least some of the nucleic acid molecules are cell-free in the biological sample, the method comprising:

for each of a plurality of the nucleic acid molecules in the biological sample:

measuring a size of the nucleic acid molecule;

identifying a location of the nucleic acid molecule in a reference genome of the organism;

identifying a first group of nucleic acid molecules as being from a first chromosomal region based on the identified locations, the first chromosomal region including a plurality of first loci;

calculating, with a computer system, a first statistical value of a size distribution of the first group of nucleic acid molecules, wherein the first statistical value is determined by computing an area under a first curve at a specified size, the first curve being a plot of cumulative frequency of nucleic acid molecules for the first chromosomal region over a range of sizes; and

comparing the first statistical value to a first reference value to determine a classification of whether the first chromosomal region exhibits an aberration, wherein:

the first reference value is a statistical value of a size distribution of a second group of nucleic acid molecules of a second chromosomal region, and

the first reference value is determined by computing an area under a second curve at the specified size, the second curve being a plot of cumulative frequency of nucleic acid molecules for the second chromosomal region over the range of sizes.

11. The method of claim 10, wherein comparing the first statistical value to the first reference value includes:

computing a difference between the first statistical value and the first reference value; and

comparing the difference to a threshold.

12. The method of claim 11, wherein the specified size is where a difference curve between the first curve and the second curve is at a maximum.
13. The method of claim 11, wherein when the difference exceeds the threshold, the classification is that the first chromosomal region potentially exhibits an aberration.
14. The method of claim 13, wherein a sign of the difference indicates a type of aberration for the first chromosomal region when the difference exceeds the threshold.
15. The method of claim 10, further comprising:
prior to calculating first statistical value, identifying the first chromosomal region as having a first aberration in the biological sample based on a first amount of nucleic acid molecules identified as being from the first chromosomal region based on the identified locations.
16. The method of claim 15, further comprising:
identifying the second chromosomal region as having a second aberration that is opposite the first aberration based on a second amount of nucleic acid molecules identified as being from the second chromosomal region based on the identified locations, wherein the classification confirms the first aberration and the second aberration.
17. The method of claim 10, wherein the range of sizes is between 0 bases and the specified size.
18. The method of claim 10, wherein the range of sizes is between 80 and the specified size.
19. The method of claim 10, wherein the specified size is between 140 bases and 190 bases.
20. A method of analyzing a biological sample of an organism, the biological sample including nucleic acid molecules originating from normal cells and potentially from cells associated with cancer, wherein at least some of the nucleic acid molecules are cell-free in the biological sample, the method comprising:

identifying a plurality of chromosomal regions of the organism, each chromosomal region including a plurality of loci, wherein the plurality of chromosomal regions are subchromosomal;

for each of a plurality of nucleic acid molecules in the biological sample of the organism:

identifying a location of the nucleic acid molecule in a reference genome of the organism;

for each of the plurality of chromosomal regions:

identifying a respective group of nucleic acid molecules as being from the chromosomal region based on the identified locations, the respective group including at least one nucleic acid molecule located at each of the plurality of loci of the chromosomal region;

calculating, with a computer system, a respective value of the respective group of nucleic acid molecules, the respective value defining a property of the nucleic acid molecules of the respective group; and

comparing the respective value to a respective reference value to determine a classification of whether the chromosomal region exhibits a deletion or an amplification; and

determining a test pattern of the chromosomal regions that exhibit a deletion or amplification, the test pattern including:

a set of the chromosomal regions that exhibit a deletion or amplification;

a first subset of the set that exhibit an amplification; and

a second subset of the set that exhibit a deletion;

comparing the test pattern to a plurality of reference patterns of different types of cancer;

based on the comparison, determining a first amount of regions of the test pattern that exhibit a same deletion or amplification as a first reference pattern corresponding to a first type of cancer; and

comparing the first amount to a first threshold to determine a first classification of whether the biological sample exhibits the first type of cancer.

21. The method of claim 20, wherein the plurality of chromosomal regions are non-overlapping.

22. The method of claim 20, wherein the property of the nucleic acid molecules of at least one of the respective groups is of one haplotype of the chromosomal region, and wherein the respective reference value is of another haplotype of the chromosomal region.

23. The method of claim 20, wherein the property of the nucleic acid molecules of at least one of the respective groups comprises a number of nucleic acid molecules with locations in the chromosomal region.

24. The method of claim 20, wherein the property of the nucleic acid molecules of at least one of the respective groups comprises a size distribution.

25. The method of claim 24, wherein the respective value comprises a mean of the size distribution, a median of the size distribution, a mode of the size distribution, or a proportion of nucleic acid molecules having a size below a size threshold.

26. The method of claim 20, wherein the plurality of reference patterns are determined from reference samples of tissues and/or mixtures of cell-free nucleic acid molecules.

27. The method of claim 20, wherein the first type of cancer is HCC, colorectal cancer, breast cancer, lung cancer, or nasopharyngeal carcinoma.

28. The method of claim 20, wherein the first reference pattern includes a first number of regions, each having a defined status of amplification, deletion, or no aberration, wherein the first amount of regions is a percentage of the first number of regions to which the test pattern matches.

29. The method of claim 20, further comprising:
determining the first threshold by:

clustering a set of reference patterns, wherein each cluster corresponds to a different type of cancer;

defining a centroid for a first cluster corresponding to the first type of cancer, wherein the centroid corresponds to regions of the reference patterns of the first cluster that are shared by at least a predetermined number of references patterns of the first cluster; and

defining a distance from a boundary of the first cluster and the centroid as the first threshold.

30. The method of claim 20, further comprising:

for each reference pattern of the plurality of reference patterns:

determining a respective amount of regions of the test pattern that exhibit a same deletion or amplification as the reference pattern; and

determining a relative likelihood for each type of cancer by comparing the respective amounts to each other; and

identifying a matching type of cancer based on the relative likelihoods.

31. A method of analyzing a biological sample of an organism, the biological sample including nucleic acid molecules originating from normal cells and from cells associated with cancer, wherein at least some of the nucleic acid molecules are cell-free in the biological sample, the method comprising:

for each of a plurality of the nucleic acid molecules in the biological sample:

measuring a size of the nucleic acid molecule;

identifying a location of the nucleic acid molecule in a reference genome of the organism;

identifying a first group of nucleic acid molecules as being from a first chromosomal region based on the identified locations, the first chromosomal region including a plurality of first loci;

calculating, with a computer system, a first statistical value of a size distribution of the first group of nucleic acid molecules;

measuring a fraction of nucleic acid molecules originating from cells associated with cancer;

selecting a first reference value based on the measured fraction;

comparing the first statistical value to the first reference value to determine a stage of cancer of the biological sample.

32. The method of claim 31, wherein selecting the first reference value based on the measured fraction includes:

selecting a long size threshold when the measured fraction is below a low cutoff, the long size threshold configured for identifying a size distribution longer than

normal, wherein an early stage of cancer is determined when the first statistical value is above the long size threshold.

33. The method of claim 31, wherein selecting the first reference value based on the measured fraction includes:

selecting a short size threshold when the measured fraction is above a high cutoff, the short size threshold configured for identifying a size distribution shorter than normal, wherein a late stage of cancer is determined when the first statistical value is below the short size threshold.

34. The method of claim 31, wherein selecting the first reference value based on the measured fraction includes:

selecting a short size threshold and a long size threshold when the measured fraction is below a high cutoff and above a low cutoff, the long size threshold configured for identifying a size distribution that is normal, wherein a middle stage of cancer is determined when the first statistical value is above the short size threshold and below the long size threshold.

35. A computer product comprising a computer readable medium storing a plurality of instructions for controlling a computer system to perform an operation of any of the methods above.

36. A system comprising:
the computer product of claim 35; and
one or more processors for executing instructions stored on the computer readable medium.

37. A system comprising means for performing any of the methods above.

38. A system configured to perform any of the above methods.

39. A system comprising modules that respectively perform the steps of any of the above methods.