# METHODS AND SYSTEMS FOR DIFFERENTIATING SOMATIC AND GERMLINE VARIANTS

## CROSS-REFERENCE

[001] This application claims the benefit of U.S. Provisional Application No. 62/561,048, filed on September 20, 2017, which application is incorporated herein by reference in its entirety.

## BACKGROUND

[002] An important aspect of cancer genomics is to precisely identify the origin of genetic alterations for appropriate treatment of patients. A recent study found that over 2% of advanced cancer patients may have unidentified germline alterations found incidentally during next-generation sequencing (NGS) for targetable somatic alterations. However, tissue-based NGS may not be able to accurately distinguish germline mutations from somatic mutations without comparison to normal tissue. In plasma, somatic variants typically occur at mutant allele fractions (MAFs) which may be 1-2 orders of magnitude lower than those of germline variants and hence liquid biopsy can accurately assign germline/somatic origin. However, certain factors such as allelic imbalance from copy number variation (CNV) or loss of heterozygosity (LOH) can skew germline MAFs away from the expected range for germline MAFs. Therefore, there is a need for methods that can take into account these factors in determining the origin of a variant.

## SUMMARY

[003] The present disclosure provides methods and systems for differentiating somatic and germline variants in a sample of nucleic acid molecules, such as cell-free deoxyribonucleic acid (cfDNA). Such methods may use common single nucleotide polymorphisms (SNPs) to model local germline allele count behavior and may distinguish somatic variants based on MAF deviation from the observed germline MAF.

[004] In one example, the present disclosure provides a method of identifying a somatic or germline origin of a nucleic acid variant from a sample of nucleic acid molecules (e.g., a tissue sample, a sample of cell-free DNA, and/or the like). The method includes (a) determining one or more quantitative measures for the nucleic acid variant from the nucleic acid sample. The quantitative measures comprise a total allele count and minor allele count for the nucleic acid variant. The method also includes (b) identifying at least one associated

variable of the nucleic acid variant from the nucleic acid sample and (c) determining a quantitative value for the associated variable of the nucleic acid variant. The method further includes (d) generating a statistical model for expected germline mutant allele counts at the genomic locus of the nucleic acid variant and (e) generating a probability value (p-value) for the nucleic acid variant based on the statistical model for expected germline allele counts and the quantitative value for the associated variable of the nucleic acid variant and at least one of the quantitative measures for the nucleic acid variant. In addition, the method also includes (f) classifying the nucleic acid variant (i) as being of somatic origin when the p-value of the nucleic acid variant is below a threshold value, or (ii) as being of germline origin when the p-value of the nucleic acid variant is at or above the threshold value.

[005] In an example, the present disclosure provides a method of identifying a somatic or germline origin of a nucleic acid variant from a sample of cell-free nucleic acid molecules (e.g., cell-free deoxyribonucleic acid (cfDNA) molecules), the method comprising: (a) determining a plurality of quantitative measures for the nucleic acid variant from the sample of cell-free nucleic acid molecules, wherein the plurality of quantitative measures comprises a total allele count and a minor allele count for the nucleic acid variant; (b) identifying an associated variable of the nucleic acid variant from the sample of cell-free nucleic acid molecules; (c) determining a quantitative value for the associated variable of the nucleic acid variant; (d) generating a statistical model for expected germline mutant allele counts at a genomic locus of the nucleic acid variant; (e) generating a probability value (p-value) for the nucleic acid variant based at least in part on the statistical model for expected germline mutant allele counts, the quantitative value for the associated variable of the nucleic acid variant, and at least one of the plurality of quantitative measures for the nucleic acid variant; and (f) classifying the nucleic acid variant as (i) being of somatic origin when the p-value for the nucleic acid variant is below a predetermined threshold value, or as (ii) being of germline origin when the p-value for the nucleic acid variant is at or above the predetermined threshold value.

[006] In some examples, the method further comprises obtaining the sample of cell-free nucleic acid molecules from a subject. In some examples, the method further comprises receiving sequencing information generated from the sample of cell-free nucleic acid molecules, wherein the sequencing information comprises cell-free nucleic acid sequencing reads comprising the nucleic acid variant and the associated variable of the

nucleic acid variant, which associated variable comprises at least one heterozygous single nucleotide polymorphism (het SNP) within a specified genomic region relative to the nucleic acid variant. In some examples, the method further comprises sequencing nucleic acids from the sample of cell-free nucleic acid molecules to generate sequencing information, wherein the plurality of quantitative measures for the nucleic acid variant and the quantitative value for the associated variable are determined from the sequencing information.

[007] In some examples, the method further comprises determining the plurality of quantitative measures for the nucleic acid variant, identifying the associated variable of the nucleic acid variant, and determining the quantitative value for the associated variable from sequencing information generated from the sample of cell-free nucleic acid molecules. In some examples, the method further comprises generating the predetermined threshold value using a beta-binomial model of expected germline mutant allele counts for nucleic acids of the sample of cell-free nucleic acid molecules. In some examples, the method further comprises classifying the somatic or germline origin of a plurality of nucleic acid variants from a plurality of genomic loci in the sample of cell-free nucleic acid molecules.

[008] In some examples, the associated variable of the nucleic acid variant comprises at least one heterozygous single nucleotide polymorphism (het SNP). In some examples, the associated variable of the nucleic acid variant comprises at least two het SNPs. In some examples, the associated variable of the nucleic acid variant comprises a genomic locus that is linked to a genomic locus that comprises the nucleic acid variant.

[009] In some examples, the method further comprises determining a mean and/or a variance value of one or more mutant allele counts for the associated variable of the nucleic acid variant. In some examples, the method further comprises determining an average quantitative value for the associated variable of the nucleic acid variant. In some examples, the associated variable of the nucleic acid variant comprises one or more of: a heterozygous single nucleotide polymorphism (het SNP), a GC content measure, a probe-specific bias measure, a fragment length value, a sequencing statistics measure, a copy number breakpoint, and clinical data for a subject. In some examples, the method further comprises determining a mean and/or a variance value of the associated variable of the nucleic acid variant.

[010] In some examples, the method further comprises determining a local germline folded mutant allele fraction (MAF), $\mu_{bin}$, for the nucleic acid variant, where *bin* is a gene or another specified genomic region comprising the nucleic acid variant and *folded MAF* is min(MAF, 1 - MAF). In some examples, the specified genomic region is a region within about $10^1$, $10^2$, $10^3$, $10^4$, $10^5$, $10^6$, $10^7$, $10^8$, $10^9$, or $10^{10}$ base pairs of the nucleic acid variant. In some examples, the associated variable of the nucleic acid variant comprises at least one single nucleotide polymorphism (SNP) comprising a population allele frequency (AF) greater than about 0.001. In some examples, the associated variable of the nucleic acid variant comprises at least one non-oncogenic single nucleotide polymorphism (SNP). In some examples, the associated variable of the nucleic acid variant comprises at least one single nucleotide polymorphism (SNP) comprising a mutant allele fraction (MAF) less than about 0.9.

[011] In some examples, the associated variable comprises at least one heterozygous single nucleotide polymorphism (SNP) within a specified genomic region relative to the nucleic acid variant, and wherein the method comprises estimating beta binomial distribution parameters using: $(x, y) \sim \text{Beta binomial}(\mu_{bin}, \rho)$, wherein y = a vector of total molecule count of the at least one germline heterozygous SNP, with one entry for each germline heterozygous SNP; x = a vector of min(mutant allele count of the at least one germline heterozygous SNP, y - mutant allele count of the at least one germline heterozygous SNP), with one entry for each germline heterozygous SNP; $\mu_{bin}$ = an estimate of the mean mutant allele count of heterozygous SNPs in a bin, wherein the bin is the specified genomic region relative to the nucleic acid variant; and ρ = an estimate of a dispersion parameter. In some examples, the method further comprises calculating an upper bound and a lower bound for the p-value. In some examples, the method further comprises calculating a two-tailed p-value for the nucleic acid variant using: $p - value = 2 * \min(Pr_{bb}(x' > A \mid \mu_{bin}, \rho, B), Pr_{bb}(x' < A \mid \mu_{bin}, \rho, B))$, wherein $Pr_{bb}$ = a probability of beta binomial; x' = a random variable distributed with the beta binomial; A = a mutant allele count of the nucleic acid variant; and B = a total molecule count of the nucleic acid variant. In some examples, ρ comprises a median value of at least one set of ρ values from a historic sample set. In some examples, the method further comprises replacing the median ρ parameter with a function of GC content of nucleic acid variant. In some examples, the method further comprises determining a maximum likelihood estimate of

$\mu_{bin}$. In some examples, the method further comprises determining a mean estimate of $\mu_{bin}$. In some examples, the method further comprises determining a maximum likelihood estimate of $\rho$. In some examples, the method further comprises determining a variance estimate of $\rho$. In some examples, the method further comprises generating a report in electronic and/or paper format with provides an indication of the classification of the nucleic acid variants as being of either somatic or germline origin.

[012] In another example, the present disclosure provides a non-transitory computer-readable medium comprising computer-executable instructions which, when executed by at least one electronic processor, perform a method comprising: (a) determining a plurality of quantitative measures for a nucleic acid variant from sequencing information generated from a sample of cell-free nucleic acid molecules (e.g., cell-free deoxyribonucleic acid (cfDNA) molecules), wherein the plurality of quantitative measures comprises a total allele count and a minor allele count for the nucleic acid variant; (b) identifying an associated variable of the nucleic acid variant from the sequencing information; (c) determining a quantitative value for the associated variable of the nucleic acid variant; (d) generating a statistical model for expected germline mutant allele counts at a genomic locus of the nucleic acid variant; (e) generating a probability value (p-value) for the nucleic acid variant based at least in part on the statistical model for expected germline mutant allele counts, the quantitative value for the associated variable of the nucleic acid variant, and at least one of the plurality of quantitative measures for the nucleic acid variant; and (f) classifying the nucleic acid variant as (i) being of somatic origin when the p-value of the nucleic acid variant is below a predetermined threshold value, or as (ii) being of germline origin when the p-value for the nucleic acid variant is at or above the predetermined threshold value.

[013] In some examples, the predetermined threshold value is generated using a beta-binomial model of expected germline mutant allele counts for the sample of cell-free nucleic acid molecules (e.g., cfDNA molecules). In some examples, the associated variable of the nucleic acid variant comprises at least one heterozygous single nucleotide polymorphism (het SNP). In some examples, the associated variable of the nucleic acid variant comprises at least two het SNPs. In some examples, the associated variable of the nucleic acid variant comprises a genomic locus that is linked to a genomic locus that comprises the nucleic acid variant. In some examples, a mean and/or a variance value of one or more mutant allele counts is determined for the associated variable of the nucleic acid

variant. In some examples, at least one of the plurality of quantitative measures comprises a number of nucleic acid molecules of the sample of cell-free nucleic acid molecules that comprise the nucleic acid variant. In some examples, the associated variable of the nucleic acid variant comprises one or more of: a heterozygous single nucleotide polymorphism (het SNP), a GC content measure, a probe-specific bias measure, a fragment length value, a sequencing statistics measure, a copy number breakpoint, and clinical data for a subject.

[014] In some examples, a local germline folded mutant allele fraction (MAF), $\mu_{bin}$, is determined for the nucleic acid variant, where *bin* is a gene or another specified genomic region comprising the nucleic acid variant and *folded MAF* is min(MAF, 1 - MAF). In some examples, the specified genomic region is a region within about $10^1$, $10^2$, $10^3$, $10^4$, $10^5$, $10^6$, $10^7$, $10^8$, $10^9$, or $10^{10}$ base pairs of the nucleic acid variant. In some examples, the associated variable of the nucleic acid variant comprises at least one single nucleotide polymorphism (SNP) comprising a population allele frequency (AF) greater than about 0.001. In some examples, the associated variable comprises at least one non-oncogenic single nucleotide polymorphism (SNP). In some examples, the associated variable of the nucleic acid variant comprises at least one single nucleotide polymorphism (SNP) comprising a mutant allele fraction (MAF) less than about 0.9.

[015] In some examples, the associated variable comprises at least one heterozygous single nucleotide polymorphism (SNP) within a specified genomic region relative to the nucleic acid variant, and wherein beta binomial distribution parameters are estimated using: $(x, y) \sim \text{Beta binomial}(\mu_{bin}, \rho)$, wherein y = a vector of total molecule count of the at least one germline heterozygous SNP, with one entry for each of the at least one germline heterozygous SNP; x = a vector of min (mutant allele count of the at least one germline heterozygous SNP, y - mutant allele count of the at least one germline heterozygous SNP), with one entry for each of the at least one germline heterozygous SNP; $\mu_{bin}$ = an estimate of the mutant allele count of heterozygous SNPs in a bin, wherein the bin is the specified genomic region relative to the nucleic acid variant; and $\rho$ = an estimate of a dispersion parameter. In some examples, an upper bound and a lower bound for the p-value are calculated. In some examples, a two-tailed p-value for the nucleic acid variant is calculated using:

$p - value = 2 * \min(Pr_{bb}(x' > x \mid \mu_{bin}, \rho, B), Pr_{bb}(x' < x \mid \mu_{bin}, \rho, B))$, wherein $Pr_{bb}$ = a probability of beta binomial; x' = a random variable distributed with the beta binomial; A = a mutant allele count of the nucleic acid variant; and B = a total molecule count of the nucleic acid variant.

[016] In another example, the present disclosure provides a system, comprising a controller comprising, or capable of accessing, a non-transitory computer-readable medium comprising computer-executable instructions which, when executed by at least one electronic processor, perform a method comprising: (a) determining a plurality of quantitative measures for a nucleic acid variant from sequencing information generated from a sample of nucleic acid molecules (e.g., a sample of cell-free deoxyribonucleic acid (cfDNA) molecules), wherein the plurality of quantitative measures comprises a total allele count and a minor allele count for the nucleic acid variant; (b) identifying an associated variable of the nucleic acid variant from the sequencing information; (c) determining a quantitative value for the associated variable of the nucleic acid variant; (d) generating a statistical model for expected germline mutant allele counts at a genomic locus of the nucleic acid variant; (e) generating a probability value (p-value) for the nucleic acid variant based at least in part on the statistical model for expected germline mutant allele counts, the quantitative value for the associated variable of the nucleic acid variant, and at least one of the plurality of quantitative measures for the nucleic acid variant; and (f) classifying the nucleic acid variant as (i) being of somatic origin when the p-value of the nucleic acid variant is below a predetermined threshold value, or as (ii) being of germline origin when the p-value for the nucleic acid variant is at or above the predetermined threshold value.

[017] In some examples, the system comprises a nucleic acid sequencer operably connected to the controller, which nucleic acid sequencer is configured to provide the sequencing information from nucleic acids of the sample of nucleic acid molecules (e.g., cell-free nucleic acid molecules). In some examples, the system comprises a sample preparation component operably connected to the controller, which sample preparation component is configured to prepare nucleic acids of the sample to be sequenced by a nucleic acid sequencer. In some examples, the system comprises a nucleic acid amplification component operably connected to the controller, which nucleic acid amplification component is configured to amplify nucleic acids of the sample. In some examples, the system comprises a material transfer component operably connected to the controller, which material

transfer component is configured to transfer one or more materials between a nucleic acid sequencer and a sample preparation component.

[018] In some examples, the predetermined threshold value is generated using a beta-binomial model of expected germline mutant allele counts for nucleic acids of the sample (e.g., cfDNA molecules). In some examples, the associated variable of the nucleic acid variant comprises at least one heterozygous single nucleotide polymorphism (het SNP). In some examples, the associated variable of the nucleic acid variant comprises at least two het SNPs. In some examples, the associated variable of the nucleic acid variant comprises a genomic locus that is linked to a genomic locus that comprises the nucleic acid variant.

[019] In some examples, a mean and/or a variance value of one or more mutant allele counts is determined for the associated variable of the nucleic acid variant. In some examples, the p-value is used to classify the nucleic acid variant. In some examples, at least one of the plurality of quantitative measures comprises a number of nucleic acid molecules of the sample of cell-free nucleic acid molecules that comprise the nucleic acid variant. In some examples, the associated variable comprises one or more of: a heterozygous single nucleotide polymorphism (het SNP), a GC content measure, a probe-specific bias measure, a fragment length value, a sequencing statistics measure, a copy number breakpoint, and clinical data for a subject.

[020] In some examples, a local germline folded mutant allele fraction (MAF), $\mu_{bin}$, is determined for the nucleic acid variant, where *bin* is a gene or another specified genomic region comprising the nucleic acid variant and *folded MAF* is min(MAF, 1 - MAF). In some examples, the specified genomic region is a region within about $10^1$, $10^2$, $10^3$, $10^4$, $10^5$, $10^6$, $10^7$, $10^8$, $10^9$, or $10^{10}$ base pairs of the nucleic acid variant. In some examples, the associated variable of the nucleic acid variant comprises at least one single nucleotide polymorphism (SNP) comprising a population allele frequency (AF) greater than about 0.001. In some examples, the associated variable of the nucleic acid variant comprises at least one non-oncogenic single nucleotide polymorphism (SNP). In some examples, the associated variable of the nucleic acid variant comprises at least one single nucleotide polymorphism (SNP) comprising a mutant allele fraction (MAF) less than about 0.9.

[021] In some examples, the associated variable comprises at least one heterozygous SNP within a specified genomic region relative to the nucleic acid variant, and wherein beta binomial distribution parameters are estimated using: $(x, y) \sim \text{Beta binomial}(\mu_{bin}, \rho)$, wherein y = a vector of total molecule count of the at least one germline heterozygous SNP, with one entry for each germline heterozygous SNP; x = a vector of min(mutant allele count of the at least one germline heterozygous SNP, y - mutant allele count of the at least one germline heterozygous SNP), with one entry for each germline heterozygous SNP; $\mu_{bin}$ = an estimate of the mutant allele count of the heterozygous SNPs in a bin, wherein the bin is the specified genomic region relative to the nucleic acid variant; and $\rho$ = an estimate of a dispersion parameter. In some examples, an upper bound and a lower bound for the p-value are calculated. In some examples, a two-tailed p-value for the nucleic acid variant is calculated using:

$p - value = 2 * \min(Pr_{bb}(x' > x \mid \mu_{bin}, \rho, B), Pr_{bb}(x' < x \mid \mu_{bin}, \rho, B))$, wherein $Pr_{bb}$ = a probability of beta binomial; x' = a random variable distributed with the beta binomial; A = a mutant allele count of the nucleic acid variant; and B = a total molecule count of the nucleic acid variant.

[022] In another aspect, the present disclosure provides a method of identifying a somatic or germline origin of a nucleic acid variant from a sample of cell-free deoxyribonucleic acid (cfDNA) molecules, the method comprising: (a) determining a mutant allele count (A) and a total molecule count (B) of the nucleic acid variant from the sample of cfDNA molecules; (b) identifying at least one germline heterozygous single nucleotide polymorphism (SNP) within a specified genomic region relative to the nucleic acid variant; (c) determining a total molecule count (y) and a mutant allele count of the at least one germline heterozygous SNP; (d) calculating a probability value (p-value) for the nucleic acid variant by: (i) determining an estimate of $\mu_{bin}$ and $\rho$ from a beta binomial distribution (x, y) ~ Beta binomial ($\mu_{bin}$, $\rho$), wherein y = a vector of total molecule count of the at least one germline heterozygous SNP, with one entry for each germline heterozygous SNP; x = a vector of min (mutant allele count of the at least one germline heterozygous SNP, y - mutant allele count of the at least one germline heterozygous SNP), with one entry for each germline heterozygous SNP; $\mu_{bin}$ = an estimate of the mutant allele count of germline heterozygous SNPs in a bin, wherein the bin is the specified genomic region relative to the nucleic acid

variant; and $\rho$ = an estimate of a dispersion parameter; (ii) calculating a two-tailed p-value from the below equation:

p-value = 2 * min($Pr_{bb}$ (x' > A | $\mu_{bin}$, $\rho$, B), $Pr_{bb}$ (x' < A | $\mu_{bin}$, $\rho$, B)), wherein $Pr_{bb}$ = a probability of beta binomial; x' = a random variable distributed with the beta binomial distribution; A = a mutant allele count of the nucleic acid variant; and B = a total molecule count of the nucleic acid variant; and (e) classifying the nucleic acid variant as (i) being of somatic origin when the p-value is below a predetermined threshold value, or as (ii) being of germline origin when the p-value is at or above the predetermined threshold value.

[023] In some examples, $\rho$ comprises a median value of at least one set of $\rho$ values from a historic sample set. In some examples, the method comprises determining a maximum likelihood estimate of $\mu_{bin}$. In some examples, the method comprises determining a mean estimate of $\mu_{bin}$. In some examples, the method comprises determining a maximum likelihood estimate of $\rho$. In some examples, the method comprises determining a variance estimate of $\rho$. In some examples, the method further comprises generating a report in electronic and/or paper format with provides an indication of the classification of the nucleic acid variants as being of either somatic or germline origin.

[024] In another example, the present disclosure provides a system, comprising a communication interface that obtains, over a communication network, sequencing information generated from nucleic acids of a sample of nucleic acid molecules (e.g., a sample of cell-free deoxyribonucleic acid (cfDNA) molecules); and a computer in communication with the communication interface, wherein the computer comprises at least one computer processor and a non-transitory computer-readable medium comprising machine-executable code that, upon execution by at least one computer processor, implements a method comprising: (a) determining a plurality of quantitative measures for a nucleic acid variant from the sequencing information, wherein the plurality of quantitative measures comprises a total allele count and a minor allele count for the nucleic acid variant; (b) identifying an associated variable of the nucleic acid variant from the sequencing information; (c) determining a quantitative value for the associated variable of the nucleic acid variant; (d) generating a statistical model for expected germline mutant allele counts at a genomic locus of the nucleic acid variant; (e) generating a probability value (p-value) for the nucleic acid variant based at least in part on the statistical model for expected germline mutant allele counts, the quantitative value for the associated variable of the nucleic acid

variant, and at least one of the plurality of quantitative measures for the nucleic acid variant; and (f) classifying the nucleic acid variant as (i) being of somatic origin when the p-value of the nucleic acid variant is below a predetermined threshold value, or as (ii) being of germline origin when the p-value for the nucleic acid variant is at or above the predetermined threshold value.

[025] In some examples, the sequencing information is provided by a nucleic acid sequencer. In some examples, the nucleic acid sequencer performs pyrosequencing, single-molecule sequencing, nanopore sequencing, semiconductor sequencing, sequencing-by-synthesis, sequencing-by-ligation, or sequencing-by-hybridization of the nucleic acids to generate the sequencing information. In some examples, the nucleic acid sequencer uses a clonal single molecule array derived from a sequencing library to generate the sequencing information. In some examples, the nucleic acid sequencer comprises a chip having an array of microwells for sequencing a sequencing library to generate the sequencing information. In some examples, the non-transitory computer-readable medium comprises a memory, a hard drive, or a memory or a hard drive of a computer server. In some examples, the communication network comprises one or more computer servers capable of distributed computing. In some examples, the distributed computing is cloud computing. In some examples, the computer is part of a computer server that is located at a location remote from the nucleic acid sequencer. In some examples, the system further comprises: an electronic display in communication with the computer over a network, wherein the electronic display comprises a user interface for displaying results upon implementing at least a portion of (a) – (f). In some examples, the user interface is a graphical user interface (GUI) or web-based user interface. In some examples, the electronic display is part of a personal computer. In some examples, the electronic display is part of an internet-enabled computer. In some examples, the internet-enabled computer is located at a location remote from the computer. In some examples, the non-transitory computer-readable medium comprises a memory, a hard drive or a memory or a hard drive of a computer server. In some examples, the communication network comprises a telecommunication network, an internet, an extranet, or an intranet.

[026] In another example, the present disclosure provides a method of treating a disease in a subject, the method comprising administering one or more customized therapies to the subject, thereby treating the disease in the subject, wherein the customized therapies have been identified by: (a) determining one or more quantitative measures for a nucleic acid

variant from a sample of nucleic acid molecules (e.g., a sample of cell-free DNA), wherein the quantitative measures comprise a total allele count and minor allele count for the nucleic acid variant; (b) identifying at least one associated variable of the nucleic acid variant from the sample of nucleic acid molecules; (c) determining a quantitative value for the associated variable of the nucleic acid variant; (d) generating a statistical model for expected germline mutant allele counts at the genomic locus of the nucleic acid variant; (e) generating a probability value (p-value) for the nucleic acid variant based on the statistical model for expected germline allele counts and the quantitative value for the associated variable of the nucleic acid variant and at least one of the quantitative measures for the nucleic acid variant; (f) classifying the nucleic acid variant (i) as being of somatic origin when the p-value of the nucleic acid variant is below a threshold value, or (ii) as being of germline origin when the p-value of the nucleic acid variant is at or above the threshold value; (g) comparing the classified nucleic acid variant to one or more comparator results that are indexed with one or more therapies; and (h) identifying one or more customized therapies for treating the disease in the subject when there is a substantial match between the classified nucleic acid variant and the comparator results.

[027] Additional examples and advantages of the present disclosure will become readily apparent to those skilled in this art from the following detailed description, wherein only illustrative examples of the present disclosure are shown and described. As will be realized, the present disclosure is capable of other and different examples, and its several details are capable of modifications in various obvious respects, all without departing from the disclosure. Accordingly, the drawings and description are to be regarded as illustrative in nature, and not as restrictive.

[028] Additional examples and advantages of the present disclosure will become readily apparent to those skilled in this art from the following detailed description, wherein only illustrative examples of the present disclosure are shown and described. As will be realized, the present disclosure is capable of other and different examples, and its several details are capable of modifications in various obvious respects, all without departing from the disclosure. Accordingly, the drawings and description are to be regarded as illustrative in nature, and not as restrictive.

## BRIEF DESCRIPTION OF THE DRAWINGS

[029] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate certain embodiments, and together with the written description,

## CLAIMS

**WHAT IS CLAIMED IS:**

1. A method of identifying a somatic or germline origin of a nucleic acid variant from a sample of cell-free deoxyribonucleic acid (cfDNA) molecules, the method comprising:

(a) determining a mutant allele count (A) and a total molecule count (B) of the nucleic acid variant from the sample of cfDNA molecules;

(b) identifying at least one germline heterozygous single nucleotide polymorphism (SNP) within a specified genomic region relative to the nucleic acid variant;

(c) determining a total molecule count (y) and a mutant allele count of the at least one germline heterozygous SNP;

(d) calculating a probability value (p-value) for the nucleic acid variant by:

(i) determining an estimate of $\mu_{bin}$ and $\rho$ from a beta binomial distribution

$$(x, y) \sim \text{Beta binomial } (\mu_{bin}, \rho),$$

wherein

y = a vector of total molecule count of the germline heterozygous SNP(s), with one entry for each germline heterozygous SNP identified in (b);

x = a vector of min (mutant allele count of the germline heterozygous SNP(s), y - mutant allele count of the germline heterozygous SNP(s)), with one entry for each germline heterozygous SNP identified in (b);

$\mu_{bin}$ = an estimate of the mutant allele count of germline heterozygous SNPs in a bin, wherein the bin is the specified genomic region relative to the nucleic acid variant; and

$\rho$ = an estimate of a dispersion parameter;

(ii) calculating a two-tailed p-value from the below equation

$$\text{p-value} = 2 * \min(Pr_{bb}(x' > A \mid \mu_{bin}, \rho, B), Pr_{bb}(x' < A \mid \mu_{bin}, \rho, B))$$

where

$Pr_{bb}$ = a probability of beta binomial;

x’ = a random variable distributed with the beta binomial distribution;

A = a mutant allele count of the nucleic acid variant;

B = a total molecule count of the nucleic acid variant; and

(e) classifying the nucleic acid variant as (i) being of somatic origin when the p-value is below a predetermined threshold value, or as (ii) being of germline origin when the p-value is at or above the predetermined threshold value.

2. The method of claim 1, wherein $\rho$ comprises a median value of at least one set of $\rho$ values from a historic sample set.

3. The method of claims 1 or 2, comprising determining a maximum likelihood estimate of $\mu_{bin}$.

4. The method of any one of claims 1 to 3, comprising determining a mean estimate of $\mu_{bin}$.

5. The method of any one of claims 1 to 4, comprising determining a maximum likelihood estimate of $\rho$.

6. The method of any one of claims 1 to 5, comprising determining a variance estimate of $\rho$.

7. The method of any one of claims 1 to 6, wherein the method further comprises generating a report in electronic and/or paper format with provides an indication of the classification of the nucleic acid variants as being of either somatic or germline origin.

8. The method of any one of claims 1 to 7, wherein the dispersion parameter is modelled as a function of GC content of the local genomic context, probe-specific bias measures, fragment length value, sequencing statistics, copy number breakpoint, and clinical data from the patient or any combination thereof.

9. The method of any one of claims 1 to 8, wherein the method comprises generating the threshold value using a beta-binomial model of expected germline mutant allele counts for nucleic acids in the sample.

10. The method of any one of claims 1 to 7 and 9, wherein the method comprises classifying the somatic or germline origin of multiple nucleic acid variants from a plurality of genomic loci in the nucleic acid sample.

11. The method of any one of claims 1 to 10, wherein the method further comprises obtaining sequence information from the cfDNA molecules from the sample from subjects by sequencing.

12. The method of any one of claims 1 to 11, wherein sequences are enriched prior to sequencing.

13. The method of any one of claims 1 to 12, wherein the cfDNA molecules are tagged with sample indexes and/or molecular barcodes.

14. The method of claim 13, wherein the cfDNA molecules are non-uniquely tagged with a limited number of barcodes such that different cfDNA molecules can be distinguished based on their endogenous sequence information in combination with at least one barcode.

15. The method of any one of claims 1 to 14, wherein the sample is plasma or serum.