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.

[003a] In one aspect, provided herein is a computer implemented 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 μ_{bin} and ρ 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);

μ_{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

ρ = 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.

[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

CLAIMS

WHAT IS CLAIMED IS:

1. A computer implemented 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 μ_{bin} and ρ 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);

μ_{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

ρ = 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 ρ comprises a median value of at least one set of ρ values from a historic sample set.

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

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

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

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

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 that 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.

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.