Claims

1. A method, implemented using a computer system comprising one or more processors and system memory, for determining a copy number variation (CNV) of a nucleic acid sequence of interest in a test sample comprising cell-free nucleic acid fragments originating from two or more genomes, the method comprising:

(a) receiving, by the computer system, sequence reads obtained by sequencing the cell-free nucleic acid fragments in the test sample;

(b) aligning, by the one or more processors, the sequence reads of the cell-free nucleic acid fragments or aligning fragments containing the sequence reads to bins of a reference genome comprising ~~a~~the sequence of interest, thereby providing test sequence tags, wherein the reference genome is divided into a plurality of bins;

(c) determining fragment sizes of at least some of the cell-free nucleic acid fragments present in the test sample;

(d) for cell-free nucleic acid fragments determined~~ing~~ as being in a first size domain, calculating, by the one or more processors, first coverages of the sequence tags for the bins of the reference genome by, for each bin:

(i) determining a number of sequence tags aligning to the bin, and

(ii) normalizing the number of sequence tags aligning to the bin by accounting for bin-to-bin variations due to factors other than copy number variation ~~using sequence tags for the cell-free nucleic acid fragments having sizes in a first size domain~~;

(e) ~~determining second coverages of the sequence tags for the bins of the reference genome using sequence tags~~ for ~~the~~ cell-free nucleic acid fragments ~~having sizes~~determined as being in a second size domain, calculating, by the one or more processors, second coverage of the sequence tags for the bins of the reference genome by, for each bin:

(i) determining a number of sequence tags aligning to the bin, and

(ii) normalizing the number of sequence tags aligning to the bin by accounting for bin-to-bin variations due to factors other than copy number variation~~wherein the second size domain is different from the first size domain~~; and

(f) determining a copy number variation in the sequence of interest using a likelihood ratio calculated from the first coverages and the second coverages.

2. The method of claim 1, wherein ~~(f) comprises determining the copy number variation in the sequence of interest using size characteristics~~the likelihood ratio is calculated from a t-statistic of the first coverages and a t-statistic of the second coverages, wherein the t-statistic is calculated using coverages of bins in the sequence of interest ~~in addition to the first coverages and the second coverages, wherein the size characteristics~~and coverages of ~~the~~ bins ~~were determined using fragment sizes of reads aligned to the bins~~in a reference region for the sequence of interest.

3. The method of claim 1, wherein the first size domain comprises cell-free nucleic acid fragments of substantially all sizes in the test sample, and the second size domain comprises only cell-free nucleic acid fragments smaller than a defined size.

4. The method of claim 1, wherein the second size domain comprises only the cell-free nucleic acid fragments smaller than about 150 bp.

~~5. The method of claim 1, wherein (f) comprises determining the copy number variation based on a likelihood ratio obtained from the first coverages and the second coverages.~~

~~6~~5. The method of claim ~~5~~1, wherein the likelihood ratio ~~comprises~~is calculated as a first likelihood that the test sample is an aneuploid sample over a second likelihood that the test sample is a euploid sample.

~~7~~6. The method of claim 1, wherein ~~(f) comprises determining the copy number variation in the sequence of interest based on a~~the likelihood ratio ~~obtained using~~is calculated from one or more values of fetal fraction~~,~~ in addition to the first coverages~~,~~ and the second coverages.

~~8~~7. The method of claim ~~7~~6, wherein the one or more values of fetal fraction comprise a value of fetal fraction calculated using ~~the fragment sizes of at least some of~~information about the sizes of the cell-free nucleic acid fragments.

~~9~~8. The method of claim ~~8~~7, wherein the value of fetal fraction is calculated by:

obtaining a frequency distribution of the ~~fragment~~ sizes of the cell-free nucleic acid fragments; and

applying the frequency distribution to a model relating fetal fraction to frequency of fragment size to obtain the fetal fraction value.

9. The method of claim 6, wherein the one or more values of fetal fraction comprise a value of fetal fraction calculated using coverage information for the bins of the reference genome.

10. The method of claim 9, wherein the value of fetal fraction is calculated by: applying coverage values of a plurality of bins to a model relating fetal fraction to coverage of bin to obtain the fetal fraction value.

11. The method of claim 6, wherein the one or more values of fetal fraction comprise a value of fetal fraction calculated using coverage information for the bins of a sex chromosome.

~~10~~12. The method of claim ~~7~~6, wherein the likelihood ratio is calculated from a fetal fraction, a t-statistic of short fragments, and a t-statistic~~s~~ of all fragments, wherein the short fragments are cell-free nucleic acid fragments in ~~the~~a first size ~~domain, which is a first size~~ range smaller than a criterion size, and the all fragments are cell-free nucleic acid fragments ~~in the second size domain, which includes~~including the short fragments and fragments longer than the criterion size.

~~11~~13. The method of claim ~~10~~12, wherein the likelihood ratio ~~comprises~~is calculated:

$$LR = \frac{\sum_{ff_{total}} q(ff_{total}) * p_1(T_{short}, T_{all}|ff_{est})}{p_0(T_{short}, T_{all})}$$

where p_1 represents ~~a~~ the likelihood that data come from a multivariate normal distribution representing a 3-copy or 1-copy model, p_0 represents ~~a~~the likelihood that data come from a multivariate normal distribution representing a 2-copy model, T_{short}, T_{all} are T scores calculated from chromosomal coverage generated from short fragments and all fragments, and $q(ff_{total})$ is a density distribution of the fetal fraction.

~~12~~14. The method of claim 1, wherein ~~the copy number variation comprises~~the likelihood ratio is calculated for monosomy X, trisomy X, trisomy 13, trisomy 18, or trisomy 21.

~~13. The method of claim 1, wherein (f) comprises obtaining a first t-statistic using the first coverages and a second t-statistic using the second coverages.~~

~~14. The method of claim 1, wherein (f) comprises determining the copy number variation in the sequence of interest using size characteristics of bins in the sequence of interest in addition to the first coverages and the second coverages, and wherein the size characteristic for a bin comprises a ratio of fragments of size smaller than a defined value to total fragments in the bin.~~

~~15. The method of claim 14, wherein (f) comprises obtaining a third t-statistic using the size characteristics of bins.~~

~~16. The method of claim 15, wherein (f) comprises calculating a likelihood ratio from a first t-statistic for the sequence of interest using the first coverages, a second t-statistic for the sequence of interest using the second coverages, and third t-statistic for the sequence of interest using the size characteristics.~~

15. The method of claim 1, wherein normalizing the number of sequence tags comprises: normalizing for GC content of the test sample, normalizing for a global wave profile of variation of a training set, and/or normalizing for one or more components obtained from a principal component analysis.

16. The method of claim 2, wherein the reference region is selected from the group consisting of: all robust chromosomes, robust chromosomes not including the sequence of interest, at least a chromosome outside of the sequence of interest, and a subset of chromosomes selected from the robust chromosomes, wherein the robust chromosomes are autosomal chromosomes other than chromosomes 13, 18, and 21.

17. The method of claim 16, wherein the reference region comprises robust chromosomes that have been determined to provide the best signal detection ability for a set of training samples.

~~17~~18. The method of claim ~~1~~2, ~~wherein (d) and/or (e) comprises~~further comprising:
calculating values of a size parameter for the bins by, for each bin:

(i) determining a ~~number of sequence tags aligning to~~value of the size parameter from sizes of cell-free nucleic acid fragments in the bin, and

(ii) normalizing the ~~number of sequence tags aligning to the bin~~value of the size parameter by accounting for bin-to-bin variations due to factors other than copy number variation; and
determining a size-based t-statistic for the sequence of interest using values of the size parameter of bins in the sequence of interest and values of the size parameter of bins in the reference region for the sequence of interest.

~~18. The method of claim 17, wherein normalizing the number of sequence tags comprises: normalizing for GC content of the sample, normalizing for a global wave profile of variation of a training set, and/or normalizing for one or more components obtained from a principal component analysis.~~

19. The method of claim 18, wherein the likelihood ratio of (f) is calculated from the first t-statistic, the second t-statistic, and the size-based t-statistic.

20. The method of claim 18, wherein the likelihood ratio of (f) is calculated from the size-based t-statistic and a fetal fraction.

21. The method of claim 1, further comprising comparing the likelihood ratio to a call criterion to determine a copy number variation in the sequence of interest.

22. The method of claim 1, further comprising obtaining a plurality of likelihood ratios and applying the plurality of likelihood ratios to a decision tree to determine a ploidy case for the test sample.

~~19~~23. A system for evaluation of copy number of a nucleic acid sequence of interest in a test sample ~~comprising cell-free nucleic acid fragments~~, the system comprising:

a sequencer for receiving cell-free nucleic acid fragments from the test sample and providing nucleic acid sequence information of the test sample;

a processor; and

one or more computer-readable storage media having stored thereon instructions for execution on said processor to:

(a) receive sequence reads obtained by sequencing the cell-free nucleic acid fragments in the test sample;

(b) align the sequence reads of the cell-free nucleic acid fragments or aligning fragments containing the sequence reads to bins of a reference genome comprising ~~a~~the sequence of interest, thereby providing test sequence tags, wherein the reference genome is divided into a plurality of bins;

(c) determine fragment sizes of at least some of the cell-free nucleic acid fragments present in the test sample;

(d) for cell-free nucleic acid fragments determined as being in a first size domain, calculate first coverages of the sequence tags for the bins of the reference genome by, for each bin:

(i) determining a number of sequence tags aligning to the bin~~using sequence tags for the cell-free nucleic acid fragments having sizes in a first size domain~~, and

(ii) normalizing the number of sequence tags aligning to the bin by accounting for bin-to-bin variations due to factors other than copy number variation;

(e) for cell-free nucleic acid fragments determined as being in a second size domain, calculate~~determine~~ second coverages of the sequence tags for the bins of the reference genome by, for each bin: ~~using sequence tags for the cell-free nucleic acid fragments having sizes in a second size domain, wherein the second size domain is different from the first size domain~~

(i) determining a number of sequence tags aligning to the bin, and

(ii) normalizing the number of sequence tags aligning to the bin by accounting for bin-to-bin variations due to factors other than copy number variation; and

(f) determine a copy number variation in the sequence of interest using a likelihood ratio calculated from the first coverages and the second coverages.

24. A method for determining a copy number variation (CNV) of a nucleic acid sequence of interest in a test sample comprising cell-free nucleic acid fragments originating from two or more genomes, the method comprising:

(a) receiving sequence reads obtained by sequencing the cell-free nucleic acid fragments in the test sample;

(b) aligning the sequence reads of the cell-free nucleic acid fragments or aligning fragments containing the sequence reads to bins of a reference genome comprising the sequence of interest, thereby providing test sequence tags, wherein the reference genome is divided into a plurality of bins;

(c) determining fragment sizes of the cell-free nucleic acid fragments existing in the test sample;

(d) calculating coverages of the sequence tags for the bins of the reference genome using sequence tags for the cell-free nucleic acid fragments having sizes in a first size domain;

(e) calculating coverages of the sequence tags for the bins of the reference genome using sequence tags for the cell-free nucleic acid fragments having sizes in a second size domain, wherein the second size domain is different from the first size domain;

(f) calculating size characteristics for the bins of the reference genome using the fragment sizes determined in (c); and

(g) determining a copy number variation in the sequence of interest using the coverages calculated in (d) and (e) and the size characteristics calculated in (f).

25. The method of claim 24, wherein (g) comprises calculating a t-statistic for the sequence of interest using the size characteristics of bins in the sequence of interest calculated in (f).

26. The method of claim 2, wherein the t-statistic is calculated as follows:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

wherein

x_1 is a bin coverage of the sequence of interest,

x_2 is a bin coverage of the reference region,

s_1 is a standard deviation of the coverages of bins in the sequence of interest,

s_2 is a standard deviation of the coverages of bins in the reference region,

n_1 is a number of the bins in the sequence of interest, and

n_2 is a number of the bins in the reference region.

27. The method of claim 1, further comprising, before (a),

extracting the cell-free nucleic acid fragments in the test sample from a plasma sample of a pregnant female carrying a fetus, wherein the cell-free nucleic acid fragments in the test sample comprise nucleic acid originating from the fetus and nucleic acid originating from the pregnant female; and

sequencing the cell-free nucleic acid fragments to obtain the sequence reads.

28. The method of claim 27, further comprising: determining that the fetus is affected by a genetic abnormality associated with the copy number variation in the sequence of interest.

29. The method of claim 28, further comprising: prescribing, initiating, and/or altering a treatment regimen, wherein the treatment regimen is designed to treat the genetic abnormality affecting the fetus.

30. The method of claim 1, further comprising, before (a),

extracting the cell-free nucleic acid fragments in the test sample from an individual, wherein the cell-free nucleic acid fragments comprise nucleic acid originating from cancer cells; and

sequencing the cell-free nucleic acid fragments to obtain the sequence reads.

31. The method of claim 30, further comprising: determining that the individual is affected by a cancer associated with the copy number variation in the sequence of interest.

32. The method of claim 31, further comprising: prescribing, initiating, and/or altering a treatment regimen, wherein the treatment regimen is designed to treat the cancer affecting the individual.

33. The method of claim 30, wherein the cell-free nucleic acid fragments in the test sample is extracted from a plasma sample of the individual.

~~20. A computer program product comprising one or more computer readable non-transitory storage media having stored thereon computer-executable instructions that, when executed by one or more processors of a computer system, cause the computer system to:~~

~~(a) receive sequence reads obtained by sequencing cell-free nucleic acid fragments in a test sample;~~

~~(b) align the sequence reads of the cell-free nucleic acid fragments or aligning fragments containing the sequence reads to bins of a reference genome comprising a sequence of interest, thereby providing test sequence tags, wherein the reference genome is divided into a plurality of bins;~~

~~(c) determine fragment sizes of at least some of the cell-free nucleic acid fragments present in the test sample;~~

~~(d) determine first coverages of the sequence tags for the bins of the reference genome using sequence tags for the cell-free nucleic acid fragments having sizes in a first size domain;~~

~~(e) determine second coverages of the sequence tags for the bins of the reference genome using sequence tags for the cell-free nucleic acid fragments having sizes in a second size domain, wherein the second size domain is different from the first size domain; and~~

~~(f) determine a copy number variation in the sequence of interest using the first coverages and the second coverages.~~