

WHAT IS CLAIMED IS:

1. An in vitro method for detecting cancer or premalignant change in a subject, the method comprising:
 - obtaining a constitutional genome of the subject;
 - 5 receiving one or more sequence tags for each of a plurality of DNA fragments in a biological sample of the subject, the biological sample including cell-free DNA;
 - determining genomic positions for the sequence tags;
 - comparing the sequence tags to the constitutional genome to determine a first number of first loci, wherein:
 - 10 at each of the first loci, a number of the sequence tags having a sequence variant relative to the constitutional genome is above a cutoff value, the cutoff value being greater than one;
 - determining a parameter based on a count of sequence tags having a sequence variant at the first loci; and
 - 15 comparing the parameter to a threshold value to determine a classification of a level of cancer in the subject.
2. The method of claim 1, wherein the threshold value is determined from one or more samples from one or more other subjects.
3. The method of claim 1, wherein the cutoff value for a locus is
 - 20 dependent on a total number of sequence tags that have a genomic position at the locus.
4. The method of claim 1, wherein different cutoff values are used for at least two of the first loci.
5. The method of claim 4, further comprising:
 - dynamically determining a first cutoff value for one of the first loci, the one of
 - 25 the first loci residing within a first region.
6. The method of claim 5, wherein the first cutoff value is determined based on a sequencing depth of the one of the first loci.

7. The method of claim 5, wherein the first cutoff value is determined based on a false positive rate that depends on a sequencing error rate, a sequencing depth of the first region, and a number of nucleotide positions in the first region.

8. The method of claim 7, wherein the first cutoff value is determined based on a number of true positives in the first region.

9. The method of claim 8, further comprising:
calculating the number of true positives for the first cutoff value based on the sequencing depth D of the first region and a fractional concentration f of tumor-derived DNA in the biological sample.

10. The method of claim 9, wherein calculating the number of true positives uses a Poisson distribution probability according to the formula: $Pb = 1 - \sum_{i=0}^{r-1} Poisson(i, M_p)$, where Pb is a probability for detecting true positives, and r is the first cutoff value, and $M_p = D \times f/2$.

11. The method of claim 5, wherein the first cutoff value is determined using any one of the following criteria:

- if a sequencing depth is less than 50 then the first cutoff value is 5,
- if the sequencing depth is 50 – 110 then the first cutoff value is 6,
- if the sequencing depth is 111 – 200 then the first cutoff value is 7,
- if the sequencing depth is 201 – 310 then the first cutoff value is 8,
- if the sequencing depth is 311 – 450 then the first cutoff value is 9,
- if the sequencing depth is 451 – 620 then the first cutoff value is 10, and
- if the sequencing depth is 621 – 800 then the first cutoff value is 11.

12. The method of claim 1, wherein the parameter is a weighted sum of the first number of first loci, wherein a contribution of each of the first loci is weighted based on an importance value assigned to the respective first loci.

13. The method of claim 1, wherein the parameter includes a sum of the sequence tags indicating a sequence variant at the first number of first loci.

14. The method of claim 13, wherein the sum is a weighted sum, and wherein one of the first loci has a first weight that is different than a second weight of a second of the first loci.

15. The method of claim 14, wherein the first weight is greater than the second weight, and wherein the one of the first loci is associated with cancer, and the second of the first loci is not associated with cancer.

16. The method of claim 1, wherein the parameter is the first number of first loci.

17. The method of claim 1, wherein determining a genomic position for a sequence tag includes:
aligning at least a portion of the sequence tags to a reference genome, wherein the alignment of a sequence tag allows for one or more mismatches between the sequence tag and the reference genome.

18. The method of claim 17, wherein comparing the sequence tags to the constitutional genome includes:

comparing the constitutional genome to the reference genome to determine a second number of second loci having a variant relative to the reference genome;

based on the aligning, determining a third number of third loci, wherein:

at each of the third loci, a number of the sequence tags having a sequence variant relative to the reference genome is above a cutoff value; and

taking a difference of the third number and the second number to obtain the first number of first loci.

19. The method of claim 18, wherein taking the difference of the third number and the second number identifies the first loci.

20. The method of claim 19, wherein determining the parameter includes: for each locus of the first number of first loci:

counting sequence tags that align to the locus and have a sequence variant at the locus; and

determining the parameter based on the respective counts.

21. The method of claim 1, wherein the constitutional genome is derived from a constitutional sample from the subject that contains more than 50% constitutional DNA.

22. The method of claim 1, wherein determining a genomic position for a sequence tag includes:

aligning at least a portion of the sequence tags to the constitutional genome, wherein the alignment of a sequence tag allows for one or more mismatches between the sequence tag and the constitutional genome.

23. The method of claim 22, wherein comparing the sequence tags to the constitutional genome includes:

based on the aligning, identifying sequence tags that have a sequence variant at a genomic location relative to the constitutional genome of the subject;

for each genomic location exhibiting a sequence variant:

counting a respective number of sequence tags that align to the genomic location and have a sequence variant at the genomic location; and

determining a parameter based on the respective numbers.

24. The method of claim 23, wherein determining the parameter based on the respective numbers includes:

summing the respective numbers to obtain a first sum; and

using the first sum to determine the parameter.

25. The method of claim 24, wherein using the first sum to determine the parameter includes:

subtracting the number of genomic locations exhibiting a sequence variant from the first sum.

26. The method of claim 24, wherein using the first sum to determine the parameter includes:

normalizing the first sum based on an amount of sequence tags aligned.

27. The method of claim 1, further comprising:

obtaining a constitutional sample of the subject that contains more than 90% constitutional DNA;

performing random sequencing of DNA fragments in the constitutional sample to obtain one or more second sequence tags for each of a plurality of DNA fragments in the constitutional sample;

aligning at least a portion of the second sequence tags to a reference genome, wherein the alignment of a second sequence tag allows for a mismatch between the second sequence tag and the reference genome at M genomic locations, wherein M is an integer equal to or greater than one; and

constructing the constitutional genome based on the second sequence tags and the aligning.

28. The method of claim 27, wherein the constitutional sample is the biological sample, and wherein constructing the constitutional genome includes:

determining a consensus sequence that includes a determination of a homozygous locus or a heterozygous locus having two alleles; and

using the consensus sequence at the constitutional genome.

29. The method of claim 1, further comprising:

receiving the biological sample of the subject; and

performing the random sequencing of DNA fragments in the biological sample to generate the one or more sequence tags for each of a plurality of DNA fragments in the biological sample.

30. The method of claim 1, wherein the one or more sequence tags are generated from a random sequencing of DNA fragments in the biological sample.

31. The method of claim 1, wherein the parameter is a fractional concentration of tumor-derived DNA.

32. A method for analyzing a heterogeneity of one or more tumors of a subject, the method comprising:

obtaining a constitutional genome of the subject;

receiving one or more sequence tags for each of a plurality of DNA fragments in a biological sample of the subject, the biological sample including cell-free DNA;

determining genomic positions for the sequence tags;

comparing the sequence tags to the constitutional genome to determine a number of first loci, wherein:

at each first loci, a first number of the sequence tags having a sequence variant relative to the constitutional genome is above a cutoff value, the cutoff value being greater than one; and

calculating a measure of heterogeneity of the one or more tumors based on the respective first numbers of the sequence tags for the first loci.

33. The method of claim 32, further comprising:

comparing the heterogeneity measure to one or more threshold values to determine a classification of a level of heterogeneity.

34. The method of claim 33, wherein the one or more threshold values are determined from one or more other subjects whose tumors have been biopsied and analyzed to determine mutations in the biopsied tumors to determine a level of heterogeneity, and wherein a measure of heterogeneity from a biological sample including cell-free DNA of one or more other subjects is used determine a threshold value.

35. The method of claim 34, wherein comparing the heterogeneity measure to one or more threshold values includes:

inputting the heterogeneity measure to a calibration function that outputs a heterogeneity level based on the heterogeneity measure.

36. The method of claim 32, wherein the heterogeneity measures includes a total number of first loci at which more than one DNA fragment aligned to the constitutional DNA while having a sequence variant at that locus.

37. The method of claim 32, wherein a plurality of heterogeneity measures are calculated, wherein calculating the heterogeneity measures includes:

for each of the first loci, calculating a proportion of sequence tags having a sequence variation;

creating a histogram of a number of first loci for values of the proportions; and identifying a plurality of peaks in the histogram.

38. The method of claim 37, wherein one of the plurality of heterogeneity measures corresponds to the number of the identified peaks.

39. The method of claim 37, wherein one of the plurality of heterogeneity measures includes a ratio of heights of two peaks of the plurality of peaks.

40. The method of claim 37, wherein the proportions each represent a fractional concentration of tumor DNA as measured at a particular first locus.

5 41. The method of claim 32, wherein the heterogeneity measure corresponds to a ratio of a first proportion of first loci having a first specified amount of sequence tags with sequence variations and a second proportion of first loci having a second specified amount of sequence tags with sequence variations.

10 42. The method of claim 41, wherein the first specified amount is less than the second specified amount.

43. The method of claim 42, wherein the first specified amount is a first range and the second specified amount is a second range, the first range being below the second range.

15 44. The method of claim 41, wherein the first specified amount and the second specified amount correspond to a fractional concentration or to an absolute number of sequence tags with a sequence variation.

20 45. The method of claim 32, wherein the one or more tumors include a plurality of clones in a single object, and wherein the heterogeneity comprises intra-tumor heterogeneity.

46. The method of claim 32, wherein the one or more tumors are a plurality of tumors, and wherein the heterogeneity comprises inter-tumor heterogeneity.

25 47. The method of claim 32, wherein the heterogeneity measure is determined from a histogram of a number of first loci corresponding to each of the respective first numbers.

48. The method of claim 47, wherein a plurality of heterogeneity measures are calculated, wherein the set of first loci include a first subset and a second subset, wherein the plurality of measures includes a first histogram of the respective first numbers corresponding to the first subset and a second histogram of the respective first numbers
5 corresponding to the second subset.

49. The method of claim 32, wherein the heterogeneity measure includes a proportion of first loci having corresponding first numbers above a specified value.

50. A method of determining a fractional concentration of tumor DNA in a biological sample including cell-free DNA, the method comprising:

10 receiving one or more sequence tags for each of a plurality of DNA fragments in the biological sample from a subject;

determining genomic positions for the sequence tags;

for each of a plurality of genomic regions that cover different parts of a reference genome:

15 determining a respective amount of DNA fragments within the genomic region from sequence tags having a genomic position within the genomic region, the respective amount of DNA fragments including DNA fragments from both chromosomal copies within the genomic region of the reference genome;

normalizing the respective amount to obtain a respective density; and

20 comparing the respective density to a reference density to identify whether the genomic region exhibits a 1-copy loss or a 1-copy gain, wherein the reference density corresponds to a density that does not exhibit a 1-copy loss or a 1-copy gain;

calculating a first density from one or more respective densities identified as exhibiting a 1-copy loss or from one or more respective densities identified as exhibiting a 1-
25 copy gain; and

calculating the fractional concentration of tumor DNA in the biological sample by:

comparing the first density to another density to obtain a differential, wherein the differential is normalized with the reference density, and wherein the another
30 density corresponds to the reference density or the one or more respective densities not used to determine the first density.

51. The method of claim 50, further comprising:

creating a histogram of a number of genomic regions having various
respective densities;

identifying a plurality of peaks in the histogram; and

calculating a heterogeneity measure from a ratio of the number of genomic
5 regions of one or more first peaks to the number of genomic regions of one or more second
peaks.

52. The method of claim 51, wherein the histogram uses values of
fractional concentrations determined separately for each genomic region in creating the
histogram.

10 53. The method of claim 51, wherein the first peaks have respective
densities of a first specified amount, and the second peaks have respective densities of a
second specified amount.

54. The method of claim 53, wherein the first specified amount is a first
range and the second specified amount is a second range, and wherein the first range is less
15 than the second range.

55. The method of claim 50, wherein comparing the respective density to
the reference density to identify whether the genomic region exhibits a 1-copy loss or a 1-
copy gain includes:

20 computing a difference between the respective density and the reference
density; and

comparing the difference to a cutoff value.

56. The method of claim 50, wherein the differential is normalized with
the reference density by:

dividing the differential by the reference density.

25 57. The method of claim 50, wherein the another density is the reference
density.

58. The method of claim 57, wherein calculating the fractional
concentration further includes multiplying the differential by two.

59. The method of claim 50, wherein the first density is calculated using respective densities identified as exhibiting a 1-copy gain, and wherein the another density is a second density calculated from respective densities identified as exhibiting a 1-copy loss.

60. The method of claim 59, wherein the differential is normalized with the reference density by:

computing a first ratio of the first density and the reference density; and
computing a second ratio of the second density and the reference density,

wherein the differential is between the first ratio and the second ratio.

61. The method of claim 59, wherein comparing the respective density to the reference density to identify whether the genomic region exhibits a 1-copy loss or a 1-copy gain includes:

fitting peaks to a distribution curve of a histogram of the respective densities, wherein the first density corresponds to a first peak and the second density corresponds to a second peak.

62. The method of claim 50, wherein all genomic regions determined to exhibit a statistically significant gain in the respective density relative to the reference density are identified as exhibiting a 1-copy gain.

63. The method of claim 48, wherein the reference density of a region is determined from a different sample than the biological sample or is determined using respective densities from other regions.

64. The method of claim 48, wherein determining a genomic position for a sequence tag includes:

aligning at least a portion of the sequence tags to a reference genome.

65. The method of claim 64, wherein the alignment of a sequence tag allows for one or more mismatches between the sequence tag and the constitutional genome.

66. The method of claim 64, wherein normalizing the respective amount to obtain the respective density of a first genomic region of the plurality of genomic regions

includes using a same total number of aligned reference tags to determine the respective density and the reference density.

67. The method of claim 64, wherein normalizing the respective amount to obtain a respective density of a first genomic region of the plurality of genomic regions
5 includes dividing the respective amount by a total number of aligned tags.

68. The method of claim 50, wherein the plurality of genomic regions each have a same length.

69. The method of claim 50, wherein the genomic regions are non-overlapping.

10 70. The method of any one of the claims above, wherein the biological sample is urine, pleural fluid, ascetic fluid, peritoneal fluid, saliva, cerebrospinal fluid, or a stool sample.

71. The method of any one of the claims above, wherein the subject is a human.

15 72. A computer program product comprising a computer readable medium storing a plurality of instructions for controlling a processor to perform an operation for methods described herein, the instructions comprising the steps of any one of the claims above.

20 73. A computer system including processors that are programmed to implement any of the methods of any of claims above.