

[0003a] In one aspect there is provided a method for detecting copy number variation comprising:

- a. attaching a set of molecular barcodes, the set of molecular barcodes comprising 2 to 1,000,000 different molecular barcodes, to a plurality of extracellular polynucleotides from a bodily fluid sample from a subject having cancer or suspected of having cancer to generate non-uniquely tagged extracellular polynucleotides;
- b. amplifying the non-uniquely tagged extracellular polynucleotides to produce amplified tagged polynucleotides;
- c. sequencing a plurality of the amplified tagged polynucleotides to generate sequence reads;
- d. mapping the sequence reads to a reference sequence;
- e. filtering out the sequence reads that fail to meet a set threshold;
- f. grouping the sequence reads into families based on sequence information from (i) molecular barcodes and (ii) at least one of a beginning (start) region of a sequence read, an end (stop) region of a sequence read, and a length of the sequence read, each family of the families corresponding to a tagged polynucleotide of the non-uniquely tagged polynucleotides;
- g. quantifying the families in two or more predefined regions of the reference sequence; and
- h. determining copy number variation in one or more of the predefined regions by:
 - i. normalizing numbers of families in the predefined regions to each other; and
 - ii. comparing the normalized numbers obtained in step (i) to normalized numbers obtained from a control sample.

[0003b] In one aspect there is provided a system comprising a computer readable medium for performing the following:

- a. mapping sequence reads generated by sequencing each of a plurality of extracellular polynucleotides from a bodily fluid sample from a subject having cancer or suspected of having cancer, wherein the extracellular polynucleotides are

non-uniquely tagged by attachment of a set of molecular barcodes, and wherein the set of molecular barcodes comprise 2 to 1,000,000 different molecular barcodes, to a reference sequence;

- b. filtering out the sequence reads that fail to meet a set threshold;
- c. grouping the sequence reads into families based on sequence information from (i) molecular barcodes and (ii) at least one of: a beginning (start) region of a sequence read, an end (stop) region of a sequence read, and a length of the sequence read, each family corresponding to a tagged polynucleotide;
- d. quantifying the families in two or more predefined regions of the reference sequence; and
- e. determining copy number variation in one or more of the predefined regions by:
 - i. normalizing numbers of families in the predefined regions to each other; and
 - ii. comparing the normalized numbers obtained in step (i) to normalized numbers obtained from a control sample.

[0003c] In one aspect there is provided a computer readable medium comprising non-transitory machine-executable code that, upon execution by a computer processor, implements a method, the method comprising:

- a. mapping sequence reads generated by sequencing each of a plurality of extracellular polynucleotides from a bodily fluid sample from a subject having cancer or suspected of having cancer, wherein the extracellular polynucleotides are non-uniquely tagged by attachment of a set of molecular barcodes, and wherein the set of molecular barcodes comprise 2 to 1,000,000 different molecular barcodes, to a reference sequence;
- b. filtering out the sequence reads that fail to meet a set threshold;
- c. grouping the sequence reads into families based on sequence information from (i) molecular barcodes and (ii) at least one of: a beginning (start) region of a sequence read, an end (stop) region of a sequence read, and a length of the sequence read, each family of the families corresponding to a tagged polynucleotide of the extracellular polynucleotides;

- d. quantifying the families in two or more predefined regions of the reference sequence;
and
- e. determining copy number variation in one or more of the predefined regions by:
 - i. normalizing numbers of families in the predefined regions to each other; and
 - ii. comparing the normalized numbers obtained in step (i) to normalized numbers obtained from a control sample.

[0004] The disclosure provides for a method for detecting copy number variation comprising: a) sequencing extracellular polynucleotides from a bodily sample from a subject, wherein each of the extracellular polynucleotide are optionally attached to unique barcodes; b) filtering out reads that fail to meet a set threshold; c) mapping sequence reads obtained from step (a) to a reference

CLAIMS

WHAT IS CLAIMED IS:

1. A method for detecting copy number variation, comprising:
 - a. attaching a set of molecular barcodes, the set of molecular barcodes comprising 2 to 1,000,000 different molecular barcodes, to a plurality of extracellular polynucleotides from a bodily fluid sample from a subject having cancer or suspected of having cancer to generate non-uniquely tagged extracellular polynucleotides;
 - b. amplifying the non-uniquely tagged extracellular polynucleotides to produce amplified tagged polynucleotides;
 - c. sequencing a plurality of the amplified tagged polynucleotides to generate sequence reads;
 - d. mapping the sequence reads to a reference sequence;
 - e. filtering out the sequence reads that fail to meet a set threshold;
 - f. grouping the sequence reads into families based on sequence information from (i) molecular barcodes and (ii) at least one of a beginning (start) region of a sequence read, an end (stop) region of a sequence read, and a length of the sequence read, each family of the families corresponding to a tagged polynucleotide of the non-uniquely tagged polynucleotides;
 - g. quantifying the families in two or more predefined regions of the reference sequence; and
 - h. determining copy number variation in one or more of the predefined regions by:
 - i. normalizing numbers of families in the predefined regions to each other; and
 - ii. comparing the normalized numbers obtained in step (i) to normalized numbers obtained from a control sample.
2. The method of claim 1, wherein the plurality of extracellular polynucleotides comprises DNA.
3. The method of claim 1, wherein the plurality of extracellular polynucleotides comprises RNA.

4. The method of claim 1, further comprising isolating the plurality of extracellular polynucleotides from the bodily fluid sample.
5. The method of claim 1, wherein the bodily fluid sample is selected from the group consisting of blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool and tears.
6. The method of claim 1, wherein the set of molecular barcodes comprises 5 to 100,000 different molecular barcodes.
7. The method of claim 6, wherein the set of molecular barcodes comprises 10 to 1,000 different molecular barcodes.
8. The method of claim 1, wherein molecular barcodes are ligated to each end of the extracellular polynucleotides in the plurality of extracellular polynucleotides.
9. The method of claim 1, further comprising selectively enriching regions from the subject's genome or transcriptome prior to sequencing.
10. The method of claim 1, further comprising non-selectively enriching regions from the subject's genome or transcriptome prior to sequencing.
11. The method of claim 6, wherein the molecular barcode is a polynucleotide.
12. The method of claim 6, wherein the molecular barcode comprises a random sequence.
13. The method of claim 6, wherein the molecular barcodes comprise oligonucleotides at least a 3, 5, 10, 15, 20, 25, 30, 35, 40, 45, or 50mer base pairs in length.
14. The method of claim 1, wherein the amplification comprises global amplification or whole genome amplification.
15. The method of claim 1, wherein the amplification comprises selective amplification.
16. The method of claim 1, wherein the amplification comprises non-selective amplification.

17. The method of claims 1, wherein suppression amplification or subtractive enrichment is performed.
18. The method of claim 1, further comprising removing a subset of the sequence reads from further analysis prior to quantifying or enumerating reads.
19. The method of claim 18, wherein removing comprises filtering out sequence reads with an accuracy or quality score of less than a threshold, e.g., 90%, 99%, 99.9%, or 99.99% and/or mapping score less than a threshold, e.g., 90%, 99%, 99.9% or 99.99%.
20. The method of claim 1, further comprising filtering out sequence reads with a quality score lower than a set threshold.
21. The method of claim 1, wherein the predefined regions are uniform or substantially uniform in size.
22. The method of claim 21, wherein the predefined regions are at least about 10kb, 20kb, 30kb 40kb, 50kb, 60kb, 70kb, 80kb, 90kb, or 100kb in size.
23. The method of claim 1, wherein at least 50, 100, 200, 500, 1000, 2000, 5000, 10,000, 20,000, or 50,000 regions are analyzed.
24. The method of claim 1, further comprising determining partial copy number variation frequency, determining loss of heterozygosity, performing gene expression analysis, performing epigenetic analysis and/or performing hypermethylation analysis.
25. The method of claim 1, wherein the normalizing and detection is performed using one or more of hidden markov, dynamic programming, support vector machine, Bayesian or probabilistic modeling, trellis decoding, Viterbi decoding, expectation maximization, Kalman filtering, or neural network methodologies.
26. The method of claim 1, further comprising monitoring disease progression, monitoring residual disease, monitoring therapy, diagnosing a condition, prognosing a condition, or selecting a therapy based on discovered variants for the subject.
27. The method of claim 26, wherein a therapy is modified based on the most recent sample analysis.

28. The method of claim 1, wherein the genetic profile of a tumor, infection or other tissue abnormality is inferred.

29. The method of claim 1, wherein the growth, remission or evolution of a tumor, infection or other tissue abnormality is monitored.

30. The method of claim 1, wherein sequences related to the subject's immune system are analyzed and monitored at single instances or over time.

31. The method of claim 1, wherein identification of a variant is followed up through an imaging test (e.g., CT, PET-CT, MRI, X-ray, ultrasound) for localization of the tissue abnormality suspected of causing the identified variant.

32. The method of claim 1, wherein the analysis further comprises use of genetic data obtained from a tissue or tumor biopsy from the same patient.

33. The method of claim 1, wherein the phylogenetics of a tumor, infection or other tissue abnormality is inferred.

34. The method of claim 1, wherein the method further comprises performing population-based no-calling and identification of low-confidence regions.

35. The method of claim 1, wherein the method further comprises obtaining measurement data for sequence coverage by measuring sequence coverage depth at every position of the genome.

36. The method of claim 35, wherein the method further comprises correcting the measurement data for sequence coverage bias by calculating window-averaged coverage.

37. The method of claim 35, wherein the method further comprises correcting the measurement data for sequence coverage bias by performing adjustments to account for GC bias in library construction and sequencing process.

38. The method of claim 35, wherein the method further comprises correcting the measurement data for sequence coverage bias by performing adjustments based on additional weighting factor associated with individual mappings to compensate for bias.

39. A system comprising a computer readable medium for performing the following:
- a. mapping sequence reads generated by sequencing each of a plurality of extracellular polynucleotides from a bodily fluid sample from a subject having cancer or suspected of having cancer, wherein the extracellular polynucleotides are non-uniquely tagged by attachment of a set of molecular barcodes, and wherein the set of molecular barcodes comprise 2 to 1,000,000 different molecular barcodes, to a reference sequence;
 - b. filtering out the sequence reads that fail to meet a set threshold;
 - c. grouping the sequence reads into families based on sequence information from (i) molecular barcodes and (ii) at least one of: a beginning (start) region of a sequence read, an end (stop) region of a sequence read, and a length of the sequence read, each family corresponding to a tagged polynucleotide;
 - d. quantifying the families in two or more predefined regions of the reference sequence; and
 - e. determining copy number variation in one or more of the predefined regions by:
 - i. normalizing numbers of families in the predefined regions to each other; and
 - ii. comparing the normalized numbers obtained in step (i) to normalized numbers obtained from a control sample.
40. The system of claim 39, wherein the entirety of the genome or at least 85% of the genome is analyzed.
41. The system of claim 39, wherein the computer readable medium provides data on percent cancer DNA or RNA in plasma or serum to the end user.
42. The method of claim 1, wherein the copy number variation identified is fractional (i.e., non-integer levels) due to heterogeneity in the sample.
43. The method of claim 1, wherein enrichment of selected regions is performed.
44. The method of claim 1, wherein copy number variation information is simultaneously extracted based on the methods described in claims 1 or 35-38.

45. The method of claim 1, further comprising an initial step of polynucleotide bottlenecking to limit the number of starting initial copies or diversity of polynucleotides in the sample.
46. The method of claim 1, further comprising the detection and/or association of affected molecular pathways.
47. The method of claim 1, further comprising serial monitoring of the health or disease state of the subject.
48. The method of claim 1, wherein the phylogeny of a genome associated with a disease within the subject is inferred.
49. The method of claim 1, wherein a disease is to be diagnosed, monitored or treated.
50. The method of claim 49, wherein the treatment regimen is selected or modified based on detected polymorphic forms or CNVs or associated pathways.
51. The method of claim 49 or 50, wherein the treatment comprises a combination therapy.
52. The method of claim 48, where the diagnosis further comprises localizing the disease using a radiographic technique, such as, a CT-Scan, PET-CT, MRI, Ultrasound, Ultrasound with microbubbles, etc.
53. A computer readable medium comprising non-transitory machine-executable code that, upon execution by a computer processor, implements a method, the method comprising:
- a. mapping sequence reads generated by sequencing each of a plurality of extracellular polynucleotides from a bodily fluid sample from a subject having cancer or suspected of having cancer, wherein the extracellular polynucleotides are non-uniquely tagged by attachment of a set of molecular barcodes, and wherein the set of molecular barcodes comprise 2 to 1,000,000 different molecular barcodes, to a reference sequence;
 - b. filtering out the sequence reads that fail to meet a set threshold;

- c. grouping the sequence reads into families based on sequence information from (i) molecular barcodes and (ii) at least one of: a beginning (start) region of a sequence read, an end (stop) region of a sequence read, and a length of the sequence read, each family of the families corresponding to a tagged polynucleotide of the extracellular polynucleotides;
- d. quantifying the families in two or more predefined regions of the reference sequence; and
- e. determining copy number variation in one or more of the predefined regions by:
 - i. normalizing numbers of families in the predefined regions to each other; and
 - ii. comparing the normalized numbers obtained in step (i) to normalized numbers obtained from a control sample.

54. The method of claim 1, wherein the plurality of extracellular polynucleotides comprises circulating tumor deoxyribonucleic acid (DNA).

55. The method of claim 1, wherein the bodily fluid sample comprises the plurality of extracellular polynucleotides in an amount of less than 1 microgram (μg).

56. The method of claim 55, wherein the amount of the plurality of extracellular polynucleotides in the bodily fluid sample is at least 10 nanograms (ng).