

SYSTEMS AND METHODS TO DETECT RARE MUTATIONS AND COPY NUMBER VARIATION

CROSS-REFERENCE

[0001] This application claims priority to U.S. Provisional Patent Application No. 61/696,734, filed September 4, 2012, U.S. Provisional Patent Application No. 61/704,400, filed September 21, 2012, U.S. Provisional Patent Application No. 61/793,997, filed March 15, 2013, and U.S. Provisional Patent Application No. 61/845,987, filed July 13, 2013, each of which is entirely incorporated herein by reference for all purposes.

BACKGROUND OF THE INVENTION

[0002] The detection and quantification of polynucleotides is important for molecular biology and medical applications such as diagnostics. Genetic testing is particularly useful for a number of diagnostic methods. For example, disorders that are caused by rare genetic alterations (e.g., sequence variants) or changes in epigenetic markers, such as cancer and partial or complete aneuploidy, may be detected or more accurately characterized with DNA sequence information.

[0003] Early detection and monitoring of genetic diseases, such as cancer is often useful and needed in the successful treatment or management of the disease. One approach may include the monitoring of a sample derived from cell free nucleic acids, a population of polynucleotides that can be found in different types of bodily fluids. In some cases, disease may be characterized or detected based on detection of genetic aberrations, such as a change in copy number variation and/or sequence variation of one or more nucleic acid sequences, or the development of other certain rare genetic alterations. Cell free DNA ("cfDNA") has been known in the art for decades, and may contain genetic aberrations associated with a particular disease. With improvements in sequencing and techniques to manipulate nucleic acids, there is a need in the art for improved methods and systems for using cell free DNA to detect and monitor disease.

SUMMARY OF THE INVENTION

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

- a. tagging each of a plurality of extracellular polynucleotides from a bodily sample from a subject with non-unique barcodes;
- b. sequencing each of the plurality of extracellular polynucleotides to generate a plurality of sequencing reads;
- c. mapping the sequence reads obtained from step (b) to a reference sequence;
- d. filtering out the sequence reads that fail to meet a set threshold;
- e. grouping the sequence reads into families based on attachment of non-unique barcodes and sequence information selected from the group consisting of: a beginning (start) region of a sequencing read, and end (stop) region of a sequencing read, and a length of the sequencing read;
- f. quantifying the families in two or more predefined regions of the reference sequence; and
- g. determining copy number variation in one or more of the predefined regions by:
 - i. normalizing number of reads in the predefined regions to each other and/or the number of unique sequence reads in the predefined regions to one another; 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 steps:

- a. mapping sequence reads generated by sequencing each of a plurality of extracellular polynucleotides from a bodily sample from a subject, said extracellular polynucleotides having been tagged with non-unique 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 attachment of non-unique barcodes and sequence information selected from the group consisting of

- d. beginning (start) region of a sequencing read, and end (stop) region of a sequencing read, and a length of the sequencing read;
- e. quantifying the families in two or more predefined regions of the reference sequence; and
- f. determining copy number variation in one or more of the predefined regions by:
 - i. normalizing number of reads in the predefined regions to each other and/or the number of unique sequence reads in the predefined regions to one another; 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 sample from a subject, said extracellular polynucleotides having been tagged with non-unique 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 attachment of non-unique barcodes and sequence information selected from the group consisting of: a

- beginning (start) region of a sequencing read, and end (stop) region of a sequencing read, and a length of the sequencing read;
- 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 number of reads in the predefined regions to each other and/or the number of unique sequence reads in the predefined regions to one another; 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. tagging each of a plurality of extracellular polynucleotides from a bodily sample from a subject with non-unique barcodes;
 - b. sequencing each of the plurality of extracellular polynucleotides to generate a plurality of sequencing reads;
 - c. mapping the sequence reads obtained from step (b) to a reference sequence;
 - d. filtering out the sequence reads that fail to meet a set threshold;
 - e. grouping the sequence reads into families based on attachment of non-unique barcodes and sequence information selected from the group consisting of: a beginning (start) region of a sequencing read, and end (stop) region of a sequencing read, and a length of the sequencing read;
 - f. quantifying the families in two or more predefined regions of the reference sequence; and
 - g. determining copy number variation in one or more of the predefined regions by:
 - i. normalizing number of reads in the predefined regions to each other and/or the number of unique sequence reads in the predefined regions to one another; 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 prevalence/concentration of each rare variant identified in the subject is reported and quantified simultaneously.
3. The method of claim 1, wherein a confidence score, regarding the prevalence/concentrations of rare variants in the subject, is reported.
4. The method of claim 1, wherein the extracellular polynucleotides comprise DNA.
5. The method of claim 1, wherein the extracellular polynucleotides comprise RNA.
6. The method of claim 1, further comprising isolating extracellular polynucleotides from the bodily sample.
7. The method claim 1, wherein the isolating comprises a method for circulating nucleic acid isolation and extraction.
8. The method of claim 1, further comprising fragmenting said isolated extracellular polynucleotides.
9. The method of claim 6 wherein the bodily sample is selected from the group consisting of blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool and tears.

10. The method of claim 1, further comprising the step of determining the percent of sequences having copy number variation or rare mutation or variant in said bodily sample.

11. The method of claim 10, wherein the determining comprises calculating the percentage of predefined regions with an amount of polynucleotides above or below a predetermined threshold.

12. The method of claim 1, wherein the subject is suspected of having an abnormal condition.

13. The method of claim 12, wherein the abnormal condition is selected from the group consisting of, mutations, rare mutations, indels, copy number variations, transversions, translocations, inversion, deletions, aneuploidy, partial aneuploidy, polyploidy, chromosomal instability, chromosomal structure alterations, gene fusions, chromosome fusions, gene truncations, gene amplification, gene duplications, chromosomal lesions, DNA lesions, abnormal changes in nucleic acid chemical modifications, abnormal changes in epigenetic patterns, abnormal changes in nucleic acid methylation infection and cancer.

14. The method of claim 1, wherein the subject is a pregnant female.

15. The method of claim 1, wherein the copy number variation or rare mutation or genetic variant is indicative of a fetal abnormality.

16. The method of claim 15, wherein the fetal abnormality is selected from the group consisting of, mutations, rare mutations, indels, copy number variations, transversions, translocations, inversion, deletions, aneuploidy, partial aneuploidy, polyploidy, chromosomal instability, chromosomal structure alterations, gene fusions, chromosome fusions, gene truncations, gene amplification, gene duplications, chromosomal lesions, DNA lesions, abnormal changes in nucleic acid chemical modifications, abnormal changes in epigenetic patterns, abnormal changes in nucleic acid methylation infection and cancer.

17. The method of claim 1, further comprising attaching one or more barcodes to the extracellular polynucleotides or fragments thereof prior to sequencing.

18. The method of claim 17, wherein each barcode attached to extracellular polynucleotides or fragments thereof prior to sequencing is unique.
19. The method of claim 17, wherein each barcode attached to extracellular polynucleotides or fragments thereof prior to sequencing is not unique.
20. The method of claim 1, further comprising selectively enriching regions from the subject's genome or transcriptome prior to sequencing.
21. The method of claim 1, further comprising non-selectively enriching regions from the subject's genome or transcriptome prior to sequencing.
22. The method of claim 1, further comprising attaching one or more barcodes to the extracellular polynucleotides or fragments thereof prior to any amplification or enrichment step.
23. The method of claim 17, wherein the barcode is a polynucleotide.
24. The method of claim 17, wherein the barcode comprises random sequence.
25. The method of claim 17, wherein the barcode comprises a fixed or semi-random set of oligonucleotides that in combination with the diversity of molecules sequenced from a select region enables identification of unique molecules.
26. The method of claim 17, wherein the barcodes comprise oligonucleotides is at least a 3, 5, 10, 15, 20, 25, 30, 35, 40, 45, or 50mer base pairs in length.
27. The method of claim 1, further comprises amplifying the extracellular polynucleotides or fragments thereof.
28. The method of claim 27, wherein the amplification comprises global amplification or whole genome amplification.

29. The method of claim 28, wherein the amplification comprises selective amplification.

30. The method of claim 29, wherein the amplification comprises non-selective amplification.

31. The method of claims 1, wherein suppression amplification or subtractive enrichment is performed.

32. The method of claim 1, further comprising removing a subset of the reads from further analysis prior to quantifying or enumerating reads.

33. The method of claim 32, wherein removing comprises filtering out 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%.

34. The method of claim 1, further comprising filtering reads with a quality score lower than a set threshold.

35. The method of claim 1, wherein the predefined regions are uniform or substantially uniform in size.

36. The method of claim 35, wherein the predefined regions are at least about 10kb, 20kb, 30kb 40kb, 50kb, 60kb, 70kb, 80kb, 90kb, or 100kb in size.

37. 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.

38. The method of claim 1, wherein the variant occurs in a region of the genome selected from the group consisting of gene fusions, gene duplications, gene deletions, gene translocations, microsatellite regions, gene fragments or combination thereof.

39. The method of claim 1, wherein the variant occurs in a region of the genome selected from the group consisting of genes, oncogenes, tumor suppressor genes, promoters, regulatory sequence elements, or combination thereof.

40. The method of claim 1, further comprising correcting/normalizing/adjusting the quantity of mapped reads using the barcodes or unique properties of individual reads.

41. The method of claim 1, wherein enumerating the reads is performed through enumeration of unique barcodes in each of the predefined regions and normalizing those numbers across at least a subset of predefined regions that were sequenced.

42. The method of claim 1, wherein samples at succeeding time intervals from the same subject are analyzed and compared to previous sample results.

43. The method of claim 40, wherein the method further comprises amplifying the barcode-attached extracellular polynucleotides.

44. 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.

45. 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.

46. 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.

47. The method of claim 46, wherein a therapy is modified based on the most recent sample analysis.

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

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

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

51. 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.

52. 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.

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

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

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

56. The method of claim 55, wherein correcting the measurement data for the sequence coverage bias comprises calculating window-averaged coverage.

57. The method of claim 55, wherein correcting the measurement data for the sequence coverage bias comprises performing adjustments to account for GC bias in the library construction and sequencing process.

58. The method of claim 55, wherein correcting the measurement data for the sequence coverage bias comprises performing adjustments based on additional weighting factor associated with individual mappings to compensate for bias.

59. The method of claim 1, wherein extracellular polynucleotide is derived from a diseased cell origin.

60. The method of claim 1, wherein extracellular polynucleotide is derived from a healthy cell origin.

61. A system comprising a computer readable medium for performing the following steps:
- a. mapping sequence reads generated by sequencing each of a plurality of extracellular polynucleotides from a bodily sample from a subject, said extracellular polynucleotides having been tagged with non-unique 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 attachment of non-unique barcodes and sequence information selected from the group consisting of beginning (start) region of a sequencing read, and end (stop) region of a sequencing read, and a length of the sequencing read;
 - 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 number of reads in the predefined regions to each other and/or the number of unique sequence reads in the predefined regions to one another; and
 - ii. comparing the normalized numbers obtained in step (i) to normalized numbers obtained from a control sample.
62. The system of claim 61, wherein the entirety of the genome or at least 85% of the genome is analyzed.
63. The system of claim 61, wherein the computer readable medium provides data on percent cancer DNA or RNA in plasma or serum to the end user.
64. The method of claim 1, whereby the copy number variants identified are fractional (i.e., non-integer levels) due to heterogeneity in the sample.
65. The method of claim 1, whereby enrichment of selected regions is performed.

66. The method of claim 1, whereby copy number variation information is simultaneously extracted based on the methods described in claims 1 or 55-58.

67. The method of claim 1, used with an initial step of polynucleotide bottlenecking to limit the number of starting initial copies or diversity of polynucleotides in the sample.

68. The method of claim 1, further comprising the detection and/or association of affected molecular pathways.

69. The method of claim 1, further comprising serial monitoring of the health or disease state of an individual.

70. The method of claim 1, whereby the phylogeny of a genome associated with a disease within an individual is inferred.

71. The method of claim 1, wherein further comprising diagnosis, monitoring or treatment of a disease.

72. The method of claim 71, wherein the treatment regimen is selected or modified based on detected polymorphic forms or CNVs or associated pathways.

73. The method of claim 71 or 72, wherein the treatment comprises a combination therapy.

74. The method of claim 70, 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.

75. 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 sample from a subject, said extracellular polynucleotides having been tagged with non-unique 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 attachment of non-unique barcodes and sequence information selected from the group consisting of: a

- beginning (start) region of a sequencing read, and end (stop) region of a sequencing read, and a length of the sequencing read;
- 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 number of reads in the predefined regions to each other and/or the number of unique sequence reads in the predefined regions to one another; and
 - ii. comparing the normalized numbers obtained in step (i) to normalized numbers obtained from a control sample.