



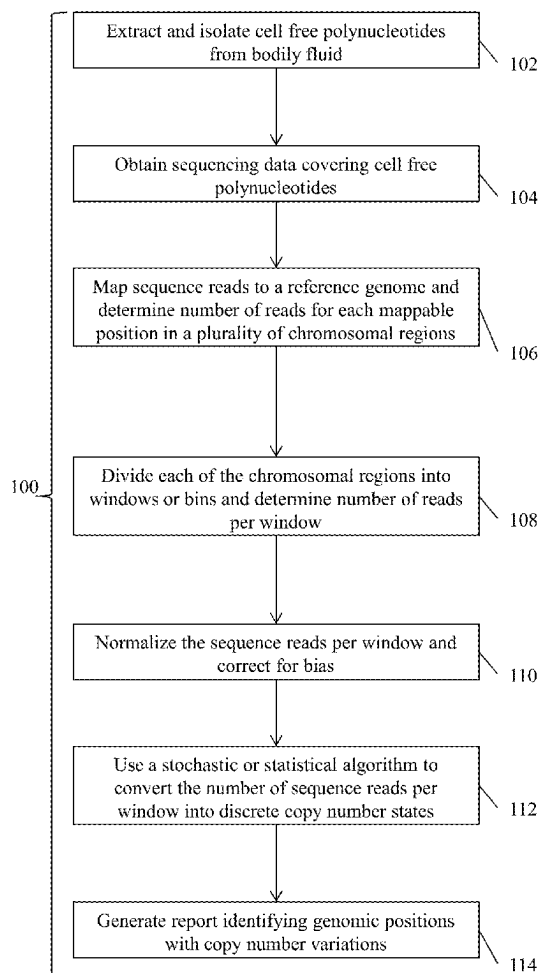
US 20140066317A1

(19) **United States**(12) **Patent Application Publication****Talasaz**(10) **Pub. No.: US 2014/0066317 A1**(43) **Pub. Date: Mar. 6, 2014**(54) **SYSTEMS AND METHODS TO DETECT
RARE MUTATIONS AND COPY NUMBER
VARIATION****Publication Classification**

(51) **Int. Cl.**
C12N 15/10 (2006.01)
(52) **U.S. Cl.**
CPC *C12N 15/1072* (2013.01)
USPC **506/2**

(71) Applicant: **Guardant Health, Inc.**, Redwood City,
CA (US)(72) Inventor: **AmirAli Talasaz**, Menlo Park, CA (US)(73) Assignee: **Guardant Health, Inc.**, Redwood City,
CA (US)(21) Appl. No.: **13/969,260**(22) Filed: **Aug. 16, 2013****Related U.S. Application Data**(60) Provisional application No. 61/696,734, filed on Sep.
4, 2012, provisional application No. 61/704,400, filed
on Sep. 21, 2012, provisional application No. 61/793,
997, filed on Mar. 15, 2013.(57) **ABSTRACT**

The present disclosure provides a system and method for the detection of rare mutations and copy number variations in cell free polynucleotides. Generally, the systems and methods comprise sample preparation, or the extraction and isolation of cell free polynucleotide sequences from a bodily fluid; subsequent sequencing of cell free polynucleotides by techniques known in the art; and application of bioinformatics tools to detect rare mutations and copy number variations as compared to a reference. The systems and methods also may contain a database or collection of different rare mutations or copy number variation profiles of different diseases, to be used as additional references in aiding detection of rare mutations, copy number variation profiling or general genetic profiling of a disease.



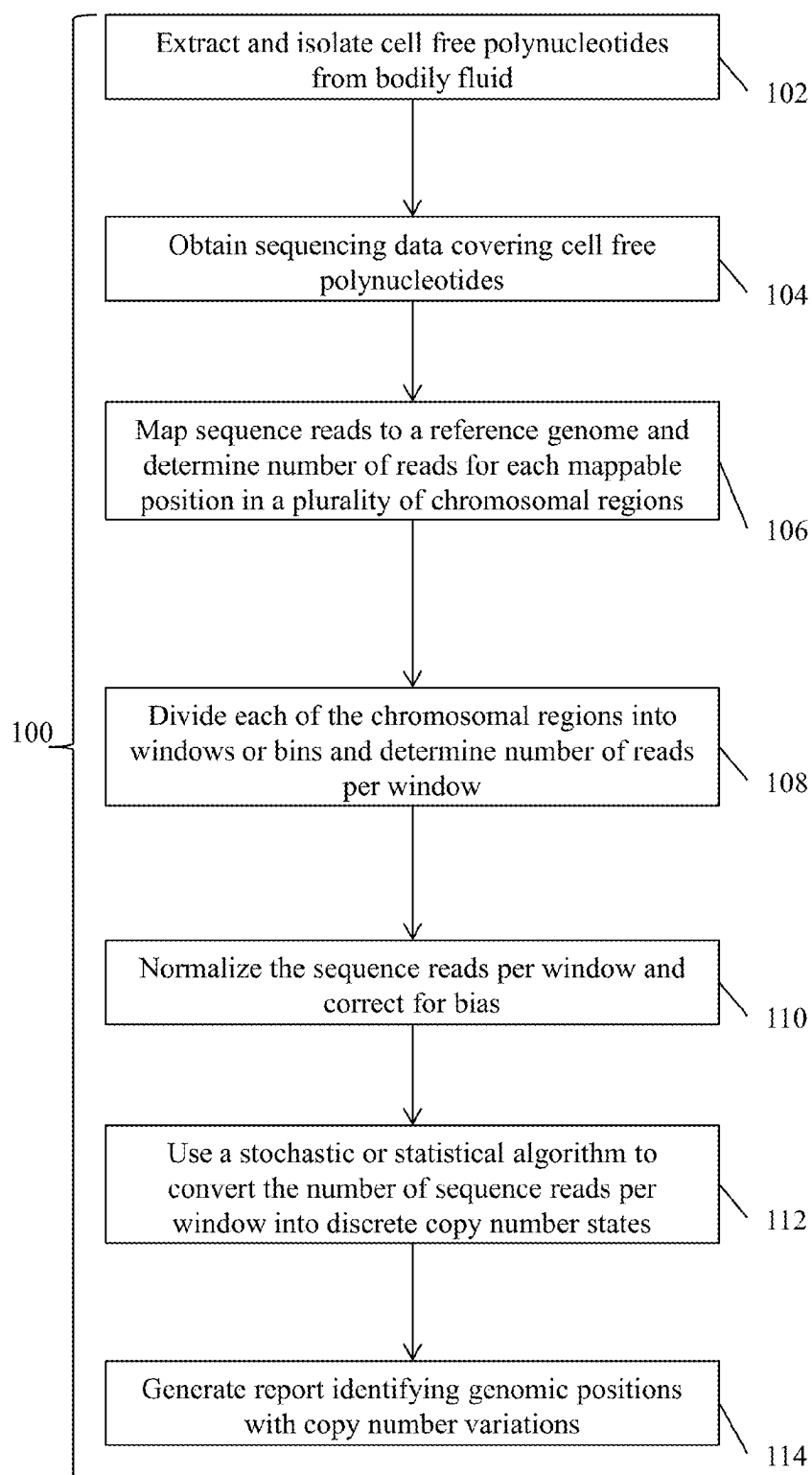


Fig. 1

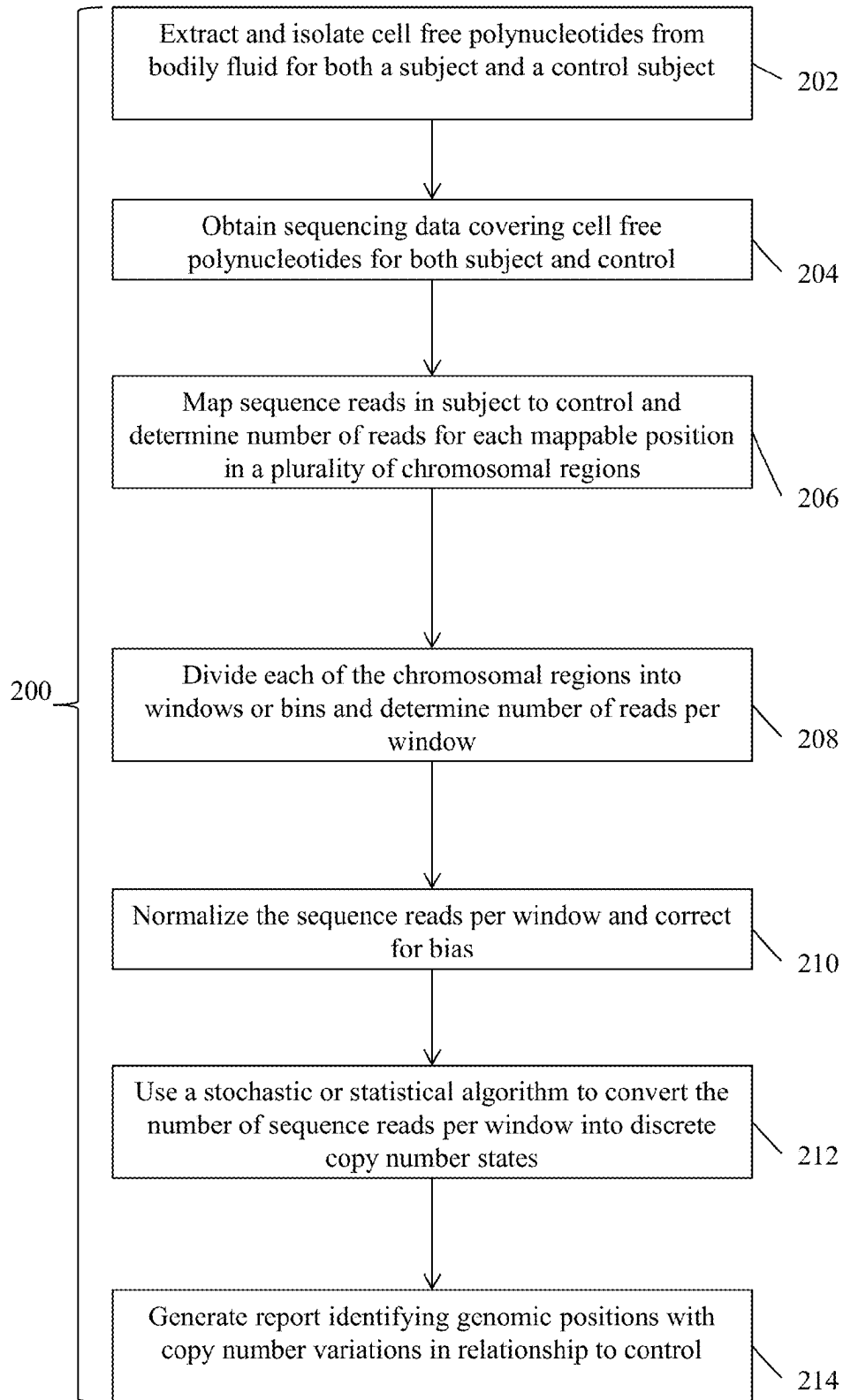


Fig. 2

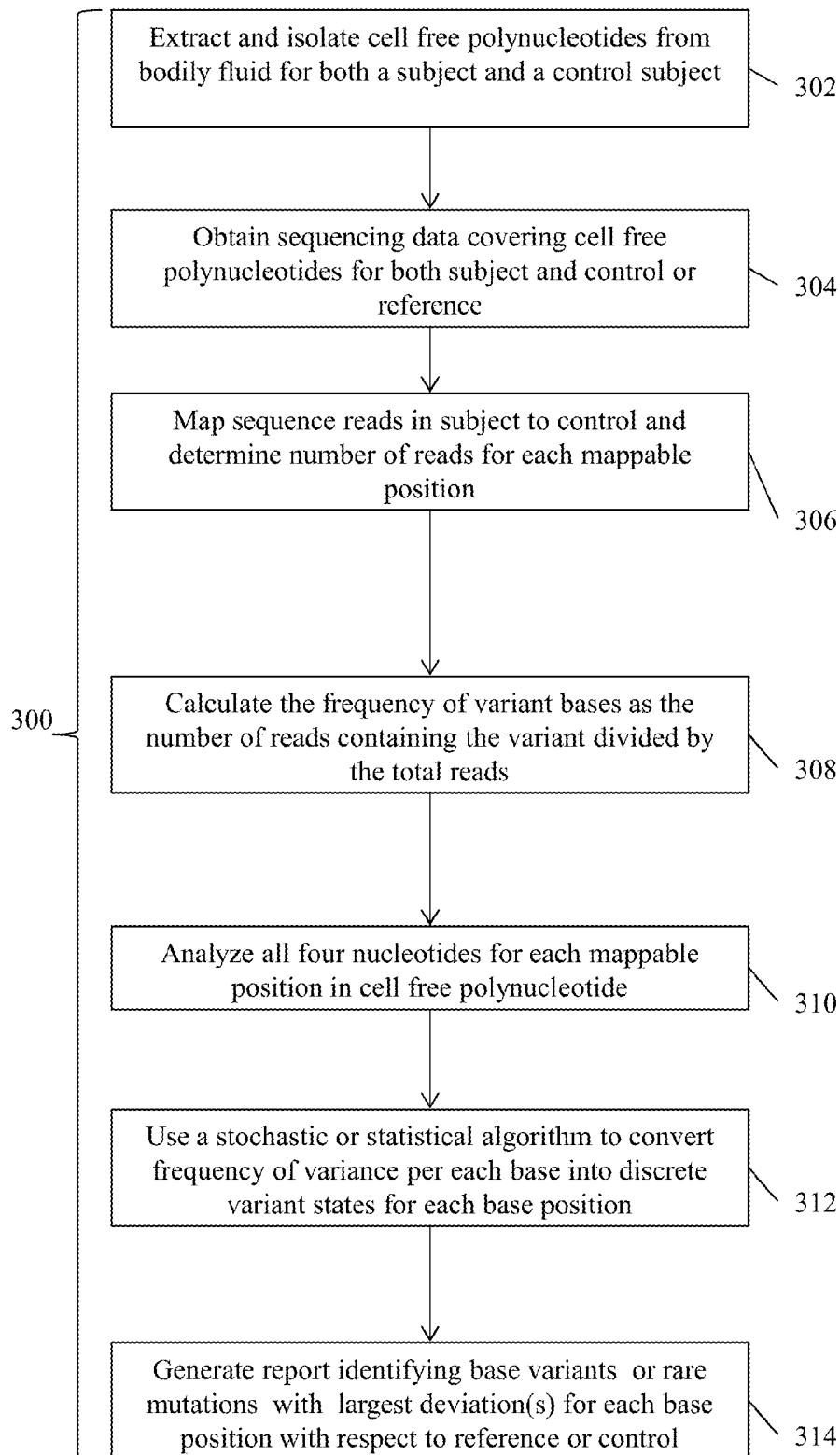


Fig. 3

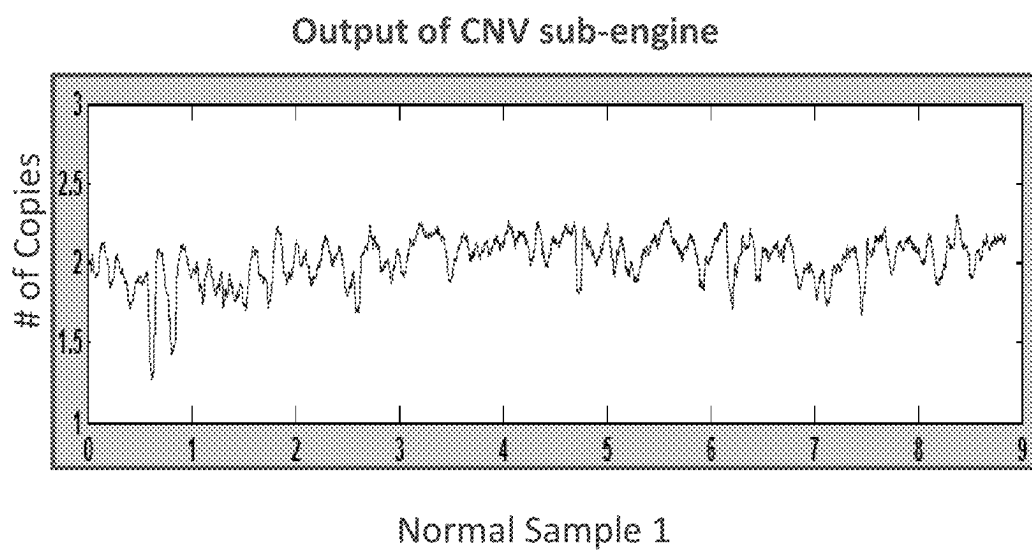


Fig. 4A

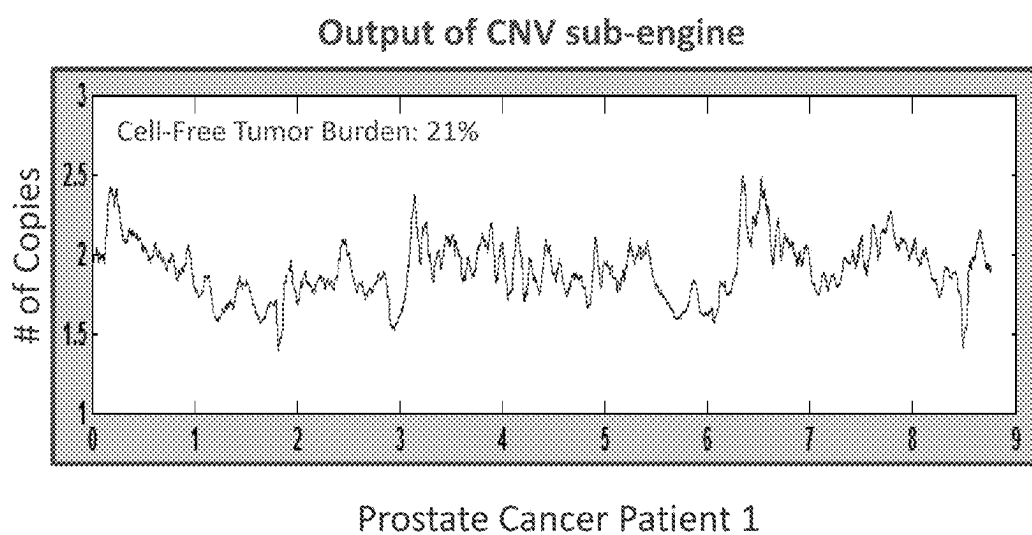


Fig. 4B

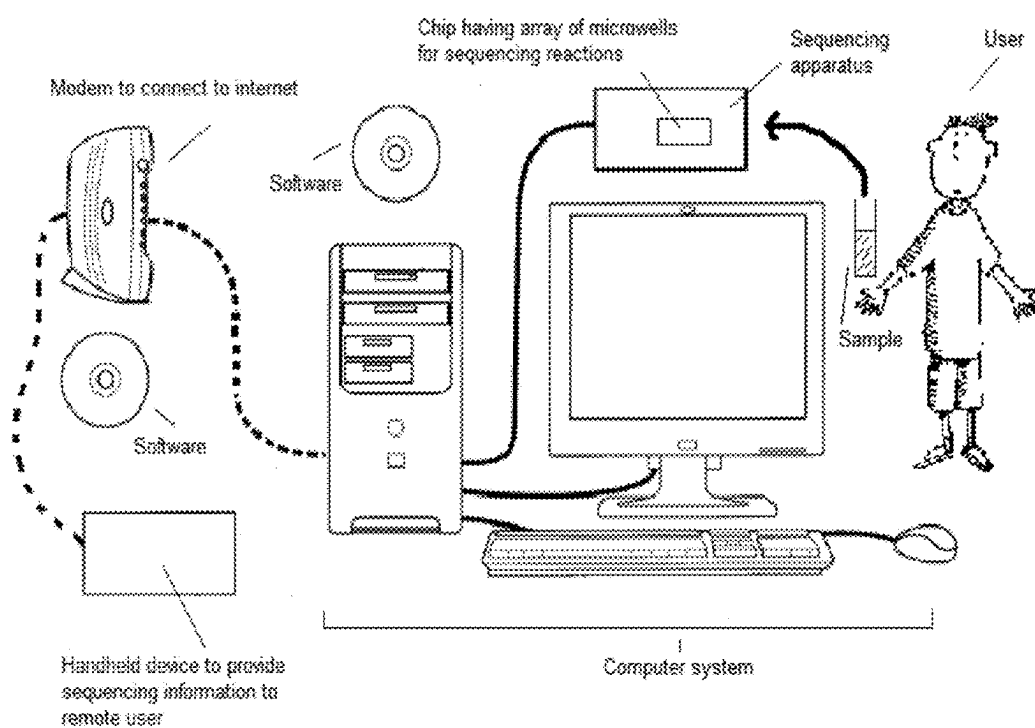


Fig. 4C

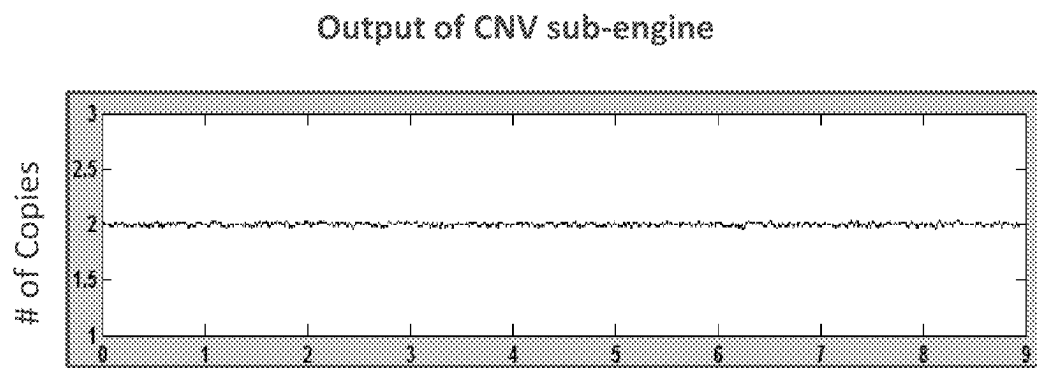


Fig. 5A Prostate Cancer Patient 2

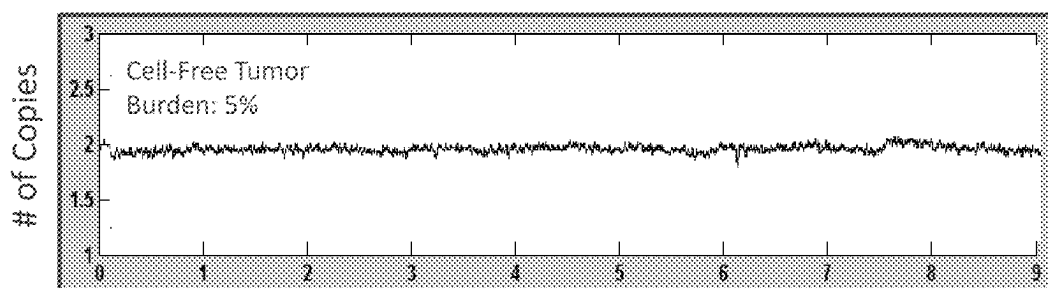


Fig. 5B Prostate Cancer Patient 3

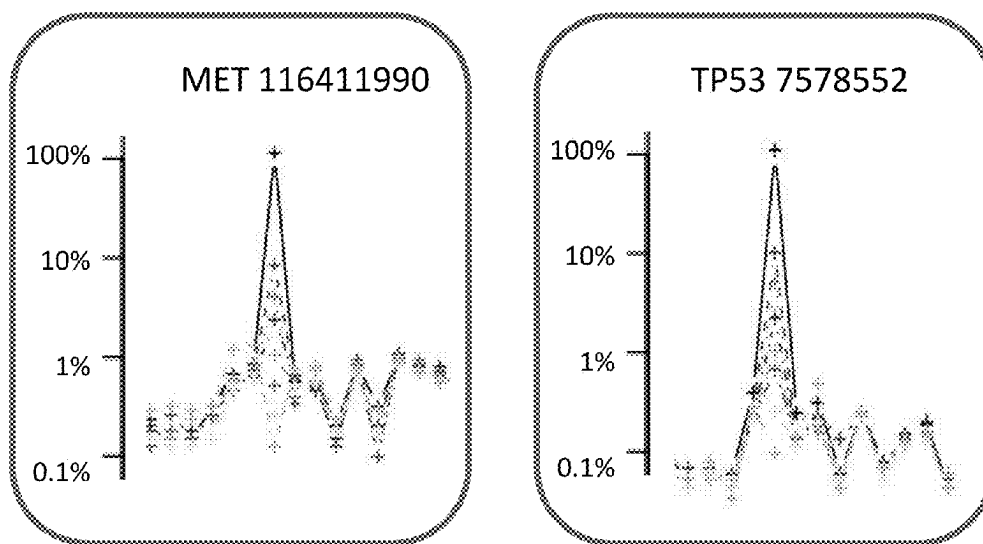


Fig. 6A

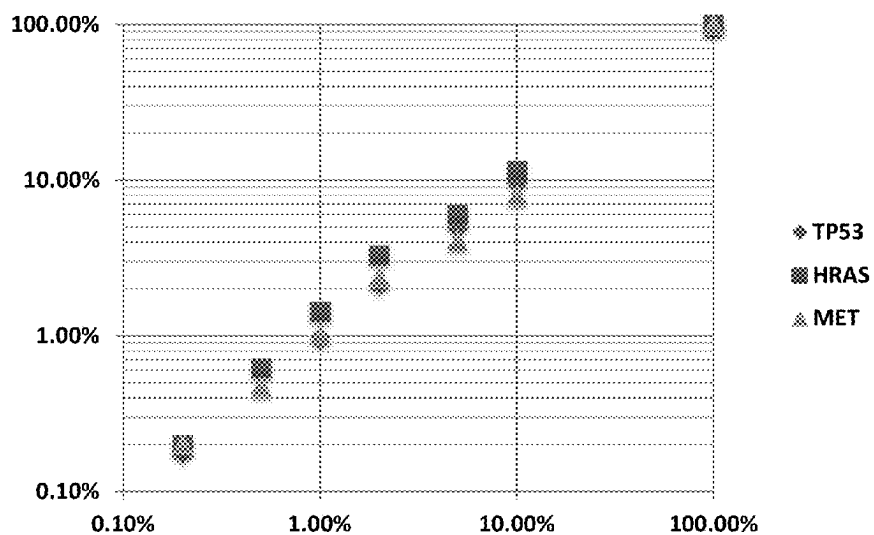


Fig. 6B

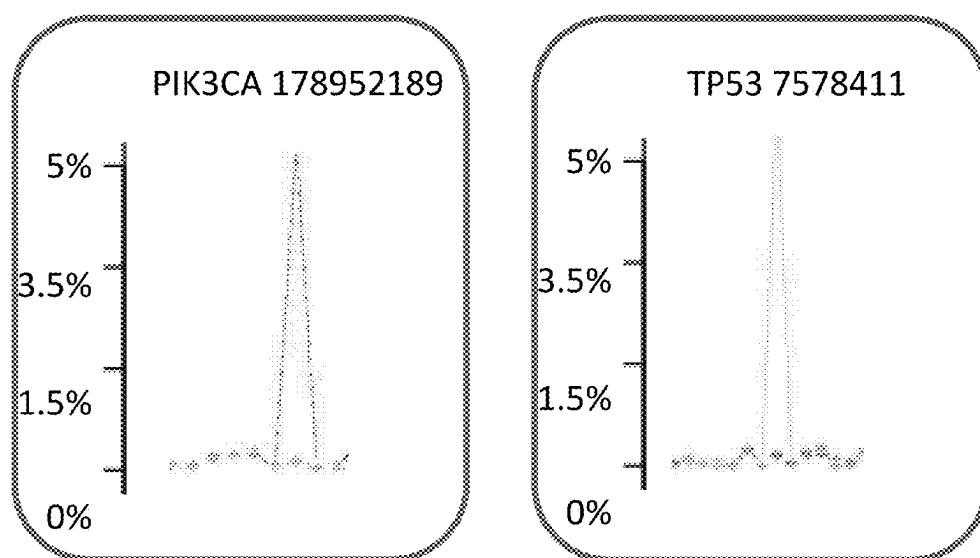


Fig. 7A

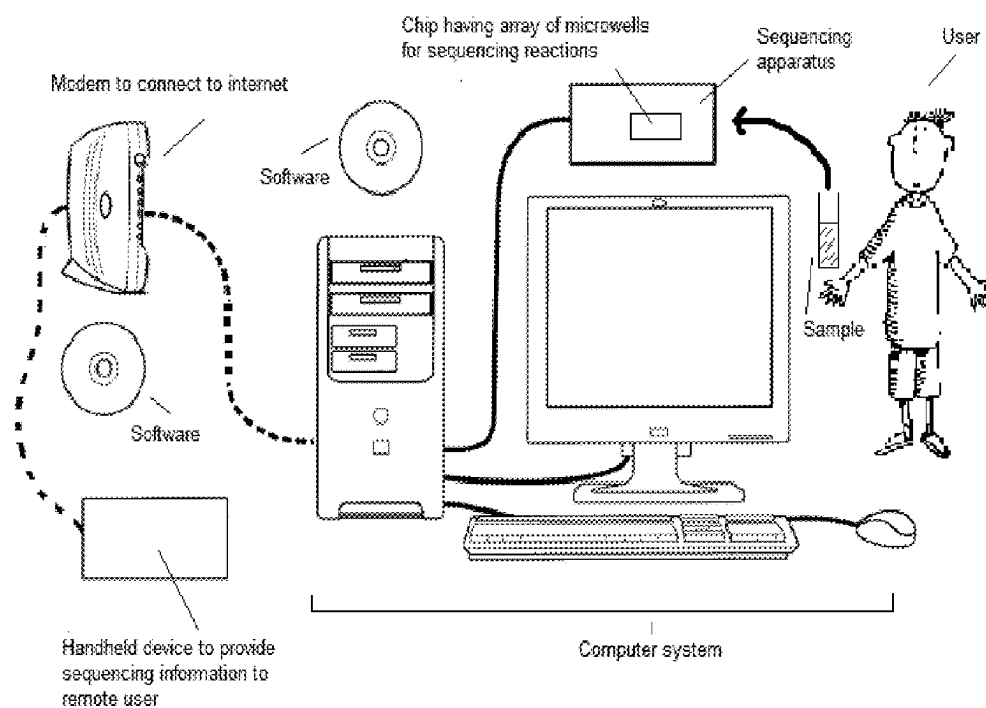


Fig. 7B

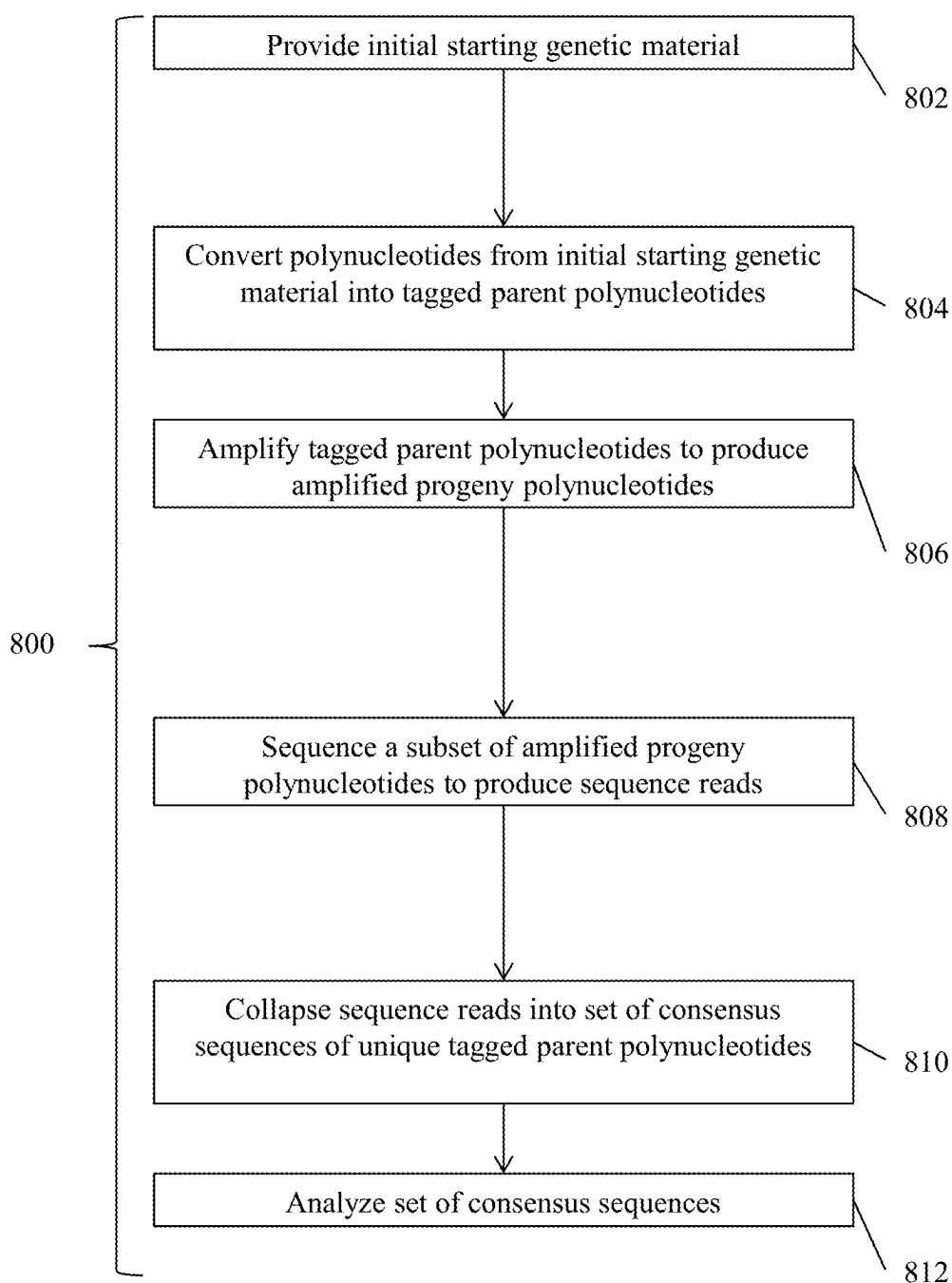


Fig. 8

SYSTEMS AND METHODS TO DETECT RARE MUTATIONS AND COPY NUMBER VARIATION

BACKGROUND OF THE INVENTION

[0001] 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 mutations, copy number variation, or changes in epigenetic markers, such as cancer and partial or complete aneuploidy, may be detected or more accurately characterized with DNA sequence information.

[0002] 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 mutation of one or more nucleic acid sequences, or the development of certain rare mutations. Cell free DNAs have 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

[0003] 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 sequence; d) quantifying/counting mapped reads in two or more predefined regions of the reference sequence; e) determining a 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 barcodes in the predefined regions to each other; (ii) comparing the normalized numbers obtained in step (i) to normalized numbers obtained from a control sample.

[0004] The disclosure also provides for a method for detecting a rare mutation in a cell-free or substantially cell free sample obtained from a subject comprising: a) sequencing extracellular polynucleotides from a bodily sample from a subject, wherein each of the extracellular polynucleotide generate a plurality of sequencing reads; sequencing extracellular polynucleotides from a bodily sample from a subject, wherein each of the extracellular polynucleotide generate a plurality of sequencing reads; b) sequencing extracellular polynucleotides from a bodily sample from a subject, wherein each of the extracellular polynucleotide generate a plurality of sequencing reads; c) filtering out reads that fail to meet a set threshold; d) mapping sequence reads derived from the sequencing onto a reference sequence; e) identifying a subset

of mapped sequence reads that align with a variant of the reference sequence at each mappable base position; f) for each mappable base position, calculating a ratio of (a) a number of mapped sequence reads that include a variant as compared to the reference sequence, to (b) a number of total sequence reads for each mappable base position; g) normalizing the ratios or frequency of variance for each mappable base position and determining potential rare variant(s) or mutation(s); h) and comparing the resulting number for each of the regions with potential rare variant(s) or mutation(s) to similarly derived numbers from a reference sample.

[0005] Additionally, the disclosure also provides for a method of characterizing the heterogeneity of an abnormal condition in a subject, the method comprising generating a genetic profile of extracellular polynucleotides in the subject, wherein the genetic profile comprises a plurality of data resulting from copy number variation and rare mutation analyses.

[0006] In some embodiments, the prevalence/concentration of each rare variant identified in the subject is reported and quantified simultaneously. In other embodiments, a confidence score, regarding the prevalence/concentrations of rare variants in the subject, is reported.

[0007] In some embodiments, extracellular polynucleotide comprises DNA. In other embodiments, extracellular polynucleotides comprise RNA. Polynucleotides may be fragments or fragmented after isolation. Additionally, the disclosure provides for a method for circulating nucleic acid isolation and extraction.

[0008] In some embodiments, extracellular polynucleotides are isolated from a bodily sample which may be selected from a group consisting of blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool and tears.

[0009] In some embodiments, the methods of the disclosure also comprise a step of determining the percent of sequences having copy number variation or rare mutation or variant in said bodily sample.

[0010] In some embodiments, the percent of sequences having copy number variation in said bodily sample is determined by calculating the percentage of predefined regions with an amount of polynucleotides above or below a predetermined threshold.

[0011] In some embodiments, bodily fluids are drawn from a subject suspected of having an abnormal condition which may be 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.

[0012] In some embodiments, the subject may be a pregnant female in which the abnormal condition may be a fetal abnormality 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

[0013] In some embodiments, the method may comprise comprising attaching one or more barcodes to the extracellular polynucleotides or fragments thereof prior to sequencing, in which the barcodes comprise are unique. In other embodiments barcodes attached to extracellular polynucleotides or fragments thereof prior to sequencing are not unique.

[0014] In some embodiments, the methods of the disclosure may comprise selectively enriching regions from the subject's genome or transcriptome prior to sequencing. In other embodiments the methods of the disclosure comprise selectively enriching regions from the subject's genome or transcriptome prior to sequencing. In other embodiments the methods of the disclosure comprise non-selectively enriching regions from the subject's genome or transcriptome prior to sequencing.

[0015] Further, the methods of the disclosure comprise attaching one or more barcodes to the extracellular polynucleotides or fragments thereof prior to any amplification or enrichment step.

[0016] In some embodiments, the barcode is a polynucleotide, which may further comprise random sequence or 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 and be at least a 3, 5, 10, 15, 20, 25, 30, 35, 40, 45, or 50mer base pairs in length.

[0017] In some embodiments, extracellular polynucleotides or fragments thereof may be amplified. In some embodiments amplification comprises global amplification or whole genome amplification.

[0018] In some embodiments, sequence reads of unique identity may be detected based on sequence information at the beginning (start) and end (stop) regions of the sequence read and the length of the sequence read. In other embodiments sequence molecules of unique identity are detected based on sequence information at the beginning (start) and end (stop) regions of the sequence read, the length of the sequence read and attachment of a barcode.

[0019] In some embodiments, amplification comprises selective amplification, non-selective amplification, suppression amplification or subtractive enrichment.

[0020] In some embodiments, the methods of the disclosure comprise removing a subset of the reads from further analysis prior to quantifying or enumerating reads.

[0021] In some embodiments, the method may comprise 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%. In other embodiments, methods of the disclosure comprise filtering reads with a quality score lower than a set threshold.

[0022] In some embodiments, predefined regions are uniform or substantially uniform in size, about 10 kb, 20 kb, 30 kb 40 kb, 50 kb, 60 kb, 70 kb, 80 kb, 90 kb, or 100 kb in size. In some embodiments, at least 50, 100, 200, 500, 1000, 2000, 5000, 10,000, 20,000, or 50,000 regions are analyzed.

[0023] In some embodiments, a genetic variant, rare mutation or copy number variation 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. In

other embodiments a genetic variant, rare mutation or copy number variation 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. In some embodiments the variant is a nucleotide variant, single base substitution, or small indel, transversion, translocation, inversion, deletion, truncation or gene truncation about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15 or 20 nucleotides in length.

[0024] In some embodiments, the method comprises correcting/normalizing/adjusting the quantity of mapped reads using the barcodes or unique properties of individual reads.

[0025] In some embodiments, 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. In some embodiments, samples at succeeding time intervals from the same subject are analyzed and compared to previous sample results. The method of the disclosure may further comprise determining partial copy number variation frequency, loss of heterozygosity, gene expression analysis, epigenetic analysis and hypermethylation analysis after amplifying the barcode-attached extracellular polynucleotides.

[0026] In some embodiments, copy number variation and rare mutation analysis is determined in a cell-free or substantially cell free sample obtained from a subject using multiplex sequencing, comprising performing over 10,000 sequencing reactions; simultaneously sequencing at least 10,000 different reads; or performing data analysis on at least 10,000 different reads across the genome. The method may comprise multiplex sequencing comprising performing data analysis on at least 10,000 different reads across the genome. The method may further comprise enumerating sequenced reads that are uniquely identifiable.

[0027] In some embodiments, the methods of the disclosure comprise normalizing and detection is performed using one or more of hidden markov, dynamic programming, support vector machine, Bayesian network, trellis decoding, Viterbi decoding, expectation maximization, Kalman filtering, or neural network methodologies.

[0028] In some embodiments the methods of the disclosure comprise monitoring disease progression, monitoring residual disease, monitoring therapy, diagnosing a condition, prognosing a condition, or selecting a therapy based on discovered variants.

[0029] In some embodiments, a therapy is modified based on the most recent sample analysis. Further, the methods of the disclosure comprise inferring the genetic profile of a tumor, infection or other tissue abnormality. In some embodiments growth, remission or evolution of a tumor, infection or other tissue abnormality is monitored. In some embodiments the subject's immune system are analyzed and monitored at single instances or over time.

[0030] In some embodiments, the methods of the disclosure comprise identification of a variant that 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.

[0031] In some embodiments, the methods of the disclosure comprise use of genetic data obtained from a tissue or tumor biopsy from the same patient. In some embodiments, whereby the phylogenetics of a tumor, infection or other tissue abnormality is inferred.

[0032] In some embodiments, the methods of the disclosure comprise performing population-based no-calling and identification of low-confidence regions. In some embodiments, obtaining the measurement data for the sequence coverage comprises measuring sequence coverage depth at every position of the genome. In some embodiments correcting the measurement data for the sequence coverage bias comprises calculating window-averaged coverage. In some embodiments correcting the measurement data for the sequence coverage bias comprises performing adjustments to account for GC bias in the library construction and sequencing process. In some embodiments 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.

[0033] In some embodiments, the methods of the disclosure comprise extracellular polynucleotide derived from a diseased cell origin. In some embodiments, the extracellular polynucleotide is derived from a healthy cell origin.

[0034] The disclosure also provides for a system comprising a computer readable medium for performing the following steps: selecting predefined regions in a genome; enumerating number of sequence reads in the predefined regions; normalizing the number of sequence reads across the predefined regions; and determining percent of copy number variation in the predefined regions. In some embodiments, the entirety of the genome or at least 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, or 90% of the genome is analyzed. In some embodiments, computer readable medium provides data on percent cancer DNA or RNA in plasma or serum to the end user.

[0035] In some embodiments, the amount of genetic variation, such as polymorphisms or causal variants is analyzed. In some embodiments, the presence or absence of genetic alterations is detected.

[0036] This disclosure also provides for a method comprising: a. providing at least one set of tagged parent polynucleotides, and for each set of tagged parent polynucleotides; b. amplifying the tagged parent polynucleotides in the set to produce a corresponding set of amplified progeny polynucleotides; c. sequencing a subset (including a proper subset) of the set of amplified progeny polynucleotides, to produce a set of sequencing reads; and d. collapsing the set of sequencing reads to generate a set of consensus sequences, each consensus sequence corresponding to a unique polynucleotide among the set of tagged parent polynucleotides. In certain embodiments the method further comprises: e. analyzing the set of consensus sequences for each set of tagged parent molecules.

[0037] In some embodiments each polynucleotide in a set is mappable to a reference sequence.

[0038] In some embodiments the method comprises providing a plurality of sets of tagged parent polynucleotides, wherein each set is mappable to a different reference sequence.

[0039] In some embodiments the method further comprises converting initial starting genetic material into the tagged parent polynucleotides.

[0040] In some embodiments the initial starting genetic material comprises no more than 100 ng of polynucleotides.

[0041] In some embodiments the method comprises bottlenecking the initial starting genetic material prior to converting.

[0042] In some embodiments the method comprises converting the initial starting genetic material into tagged parent polynucleotides with a conversion efficiency of at least 10%, at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 80% or at least 90%.

[0043] In some embodiments converting comprises any of blunt-end ligation, sticky end ligation, molecular inversion probes, PCR, ligation-based PCR, single strand ligation and single strand circularization.

[0044] In some embodiments the initial starting genetic material is cell-free nucleic acid.

[0045] In some embodiments a plurality of the reference sequences are from the same genome.

[0046] In some embodiments each tagged parent polynucleotide in the set is uniquely tagged.

[0047] In some embodiments the tags are non-unique.

[0048] In some embodiments the generation of consensus sequences is based on information from the tag and at least one of sequence information at the beginning (start) region of the sequence read, the end (stop) regions of the sequence read and the length of the sequence read.

[0049] In some embodiments the method comprises sequencing a subset of the set of amplified progeny polynucleotides sufficient to produce sequence reads for at least one progeny from each of at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 80%, at least 90% at least 95%, at least 98%, at least 99%, at least 99.9% or at least 99.99% of unique polynucleotides in the set of tagged parent polynucleotides.

[0050] In some embodiments the at least one progeny is a plurality of progeny, e.g., at least 2, at least 5 or at least 10 progeny.

[0051] In some embodiments the number of sequence reads in the set of sequence reads is greater than the number of unique tagged parent polynucleotides in the set of tagged parent polynucleotides.

[0052] In some embodiments the subset of the set of amplified progeny polynucleotides sequenced is of sufficient size so that any nucleotide sequence represented in the set of tagged parent polynucleotides at a percentage that is the same as the percentage per-base sequencing error rate of the sequencing platform used, has at least a 50%, at least a 60%, at least a 70%, at least a 80%, at least a 90% at least a 95%, at least a 98%, at least a 99%, at least a 99.9% or at least a 99.99% chance of being represented among the set of consensus sequences.

[0053] In some embodiments the method comprises enriching the set of amplified progeny polynucleotides for polynucleotides mapping to one or more selected reference sequences by: (i) selective amplification of sequences from initial starting genetic material converted to tagged parent polynucleotides; (ii) selective amplification of tagged parent polynucleotides; (iii) selective sequence capture of amplified progeny polynucleotides; or (iv) selective sequence capture of initial starting genetic material.

[0054] In some embodiments analyzing comprises normalizing a measure (e.g., number) taken from a set of consensus sequences against a measure taken from a set of consensus sequences from a control sample.

[0055] In some embodiments analyzing comprises detecting 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 or cancer.

[0056] In some embodiments the polynucleotides comprise DNA, RNA, a combination of the two or DNA plus RNA-derived cDNA.

[0057] In some embodiments a certain subset of polynucleotides is selected for or is enriched based on polynucleotide length in base-pairs from the initial set of polynucleotides or from the amplified polynucleotides.

[0058] In some embodiments analysis further comprises detection and monitoring of an abnormality or disease within an individual, such as, infection and/or cancer.

[0059] In some embodiments the method is performed in combination with immune repertoire profiling.

[0060] In some embodiments the polynucleotides are extract from the group consisting of blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool, and tears.

[0061] In some embodiments collapsing comprising detecting and/or correcting errors, nicks or lesions present in the sense or anti-sense strand of the tagged parent polynucleotides or amplified progeny polynucleotides.

[0062] This disclosure also provides for a method comprising detecting genetic variation in initial starting genetic material with a sensitivity of at least 5%, at least 1%, at least 0.5%, at least 0.1% or at least 0.05%. In some embodiments the initial starting genetic material is provided in an amount less than 100 ng of nucleic acid, the genetic variation is copy number/heterozygosity variation and detecting is performed with sub-chromosomal resolution; e.g., at least 100 megabase resolution, at least 10 megabase resolution, at least 1 megabase resolution, at least 100 kilobase resolution, at least 10 kilobase resolution or at least 1 kilobase resolution.

[0063] This disclosure also provides for a system comprising a computer readable medium for performing the following steps: a. providing at least one set of tagged parent polynucleotides, and for each set of tagged parent polynucleotides; b. amplifying the tagged parent polynucleotides in the set to produce a corresponding set of amplified progeny polynucleotides; c. sequencing a subset (including a proper subset) of the set of amplified progeny polynucleotides, to produce a set of sequencing reads; and d. collapsing the set of sequencing reads to generate a set of consensus sequences, each consensus sequence corresponding to a unique polynucleotide among the set of tagged parent polynucleotides and, optionally, e. analyzing the set of consensus sequences for each set of tagged parent molecules.

BRIEF DESCRIPTION OF THE DRAWINGS

[0064] The novel features of a system and methods of this disclosure are set forth with particularity in the appended claims. A better understanding of the features and advantages of this disclosure will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of a systems and methods of this disclosure are utilized, and the accompanying drawings of which:

[0065] FIG. 1 is a flow chart representation of a method of detection of copy number variation using a single sample.

[0066] FIG. 2 is a flow chart representation of a method of detection of copy number variation using paired samples.

[0067] FIG. 3 is a flow chart representation of a method of detection of rare mutation detection.

[0068] FIG. 4A is graphical copy number variation detection report generated from a normal, non cancerous subject.

[0069] FIG. 4B is a graphical copy number variation detection report generated from a subject with prostate cancer.

[0070] FIG. 4C is schematic representation of internet enabled access of reports generated from copy number variation analysis of a subject with prostate cancer.

[0071] FIG. 5A is a graphical copy number variation detection report generated from a subject with prostate cancer remission.

[0072] FIG. 5B is a graphical copy number variation detection report generated from a subject with prostate recurrence cancer.

[0073] FIG. 6A is graphical rare mutation detection report generated from various mixing experiments using DNA samples containing both wildtype and mutant copies of MET and TP53.

[0074] FIG. 6B is logarithmic graphical representation of rare mutation detection results. Observed vs. expected percent cancer measurements are shown for various mixing experiments using DNAs samples containing both wildtype and mutant copies of MET, HRAS and TP53.

[0075] FIG. 7A is graphical report of percentage of two rare mutations in two genes, MET and TP53, in a subject with prostate cancer as compared to a reference (control).

[0076] FIG. 7B is schematic representation of internet enabled access of reports generated from rare mutation analysis of a subject with prostate cancer.

[0077] FIG. 8 is a flow chart representation of a method of analyzing genetic material.

DETAILED DESCRIPTION OF THE INVENTION

I. General Overview

[0078] The present disclosure provides a system and method for the detection of rare mutations and copy number variations in cell free polynucleotides. Generally, the systems and methods comprise sample preparation, or the extraction and isolation of cell free polynucleotide sequences from a bodily fluid; subsequent sequencing of cell free polynucleotides by techniques known in the art; and application of bioinformatics tools to detect rare mutations and copy number variations as compared to a reference. The systems and methods also may contain a database or collection of different rare mutations or copy number variation profiles of different diseases, to be used as additional references in aiding detection of rare mutations, copy number variation profiling or general genetic profiling of a disease.

[0079] The systems and methods may be particularly useful in the analysis of cell free DNAs. In some cases, cell free DNAs are extracted and isolated from a readily accessible bodily fluid such as blood. For example, cell free DNAs can be extracted using a variety of methods known in the art, including but not limited to isopropanol precipitation and/or silica based purification. Cell free DNAs may be extracted from any number of subjects, such as subjects without cancer, subjects at risk for cancer, or subjects known to have cancer (e.g. through other means).

[0080] Following the isolation/extraction step, any of a number of different sequencing operations may be performed on the cell free polynucleotide sample. Samples may be processed before sequencing with one or more reagents (e.g.,

enzymes, unique identifiers (e.g., barcodes), probes, etc.). In some cases if the sample is processed with a unique identifier such as a barcode, the samples or fragments of samples may be tagged individually or in subgroups with the unique identifier. The tagged sample may then be used in a downstream application such as a sequencing reaction by which individual molecules may be tracked to parent molecules.

[0081] After sequencing data of cell free polynucleotide sequences is collected, one or more bioinformatics processes may be applied to the sequence data to detect genetic features or aberrations such as copy number variation, rare mutations or changes in epigenetic markers, including but not limited to methylation profiles. In some cases, in which copy number variation analysis is desired, sequence data may be: 1) aligned with a reference genome; 2) filtered and mapped; 3) partitioned into windows or bins of sequence; 4) coverage reads counted for each window; 5) coverage reads can then be normalized using a stochastic or statistical modeling algorithm; 6) and an output file can be generated reflecting discrete copy number states at various positions in the genome. In other cases, in which rare mutation analysis is desired, sequence data may be 1) aligned with a reference genome; 2) filtered and mapped; 3) frequency of variant bases calculated based on coverage reads for that specific base; 4) variant base frequency normalized using a stochastic, statistical or probabilistic modeling algorithm; 5) and an output file can be generated reflecting mutation states at various positions in the genome.

[0082] A variety of different reactions and/or operations may occur within the systems and methods disclosed herein, including but not limited to: nucleic acid sequencing, nucleic acid quantification, sequencing optimization, detecting gene expression, quantifying gene expression, genomic profiling, cancer profiling, or analysis of expressed markers. Moreover, the systems and methods have numerous medical applications. For example, it may be used for the identification, detection, diagnosis, treatment, staging of, or risk prediction of various genetic and non-genetic diseases and disorders including cancer. It may be used to assess subject response to different treatments of said genetic and non-genetic diseases, or provide information regarding disease progression and prognosis.

[0083] The present disclosure further provides methods and systems for detecting with high sensitivity genetic variation in a sample of initial genetic material. The methods involve using one or both of the following tools: First, the efficient conversion of individual polynucleotides in a sample of initial genetic material into sequence-ready tagged parent polynucleotides, so as to increase the probability that individual polynucleotides in a sample of initial genetic material will be represented in a sequence-ready sample. This can produce sequence information about more polynucleotides in the initial sample. Second, high yield generation of consensus sequences for tagged parent polynucleotides by high rate sampling of progeny polynucleotides amplified from the tagged parent polynucleotides, and collapsing of generated sequence reads into consensus sequences representing sequences of parent tagged polynucleotides. This can reduce noise introduced by amplification bias and/or sequencing errors, and can increase sensitivity of detection.

[0084] Sequencing methods typically involve sample preparation, sequencing of polynucleotides in the prepared sample to produce sequence reads and bioinformatic manipulation of the sequence reads to produce quantitative and/or

qualitative genetic information about the sample. Sample preparation typically involves converting polynucleotides in a sample into a form compatible with the sequencing platform used. This conversion can involve tagging polynucleotides. In certain embodiments of this invention the tags comprise polynucleotide sequence tags. Conversion methodologies used in sequencing may not be 100% efficient. For example, it is not uncommon to convert polynucleotides in a sample with a conversion efficiency of about 1-5%, that is, about 1-5% of the polynucleotides in a sample are converted into tagged polynucleotides. Polynucleotides that are not converted into tagged molecules are not represented in a tagged library for sequencing. Accordingly, polynucleotides having genetic variants represented at low frequency in the initial genetic material may not be represented in the tagged library and, therefore may not be sequenced or detected. By increasing conversion efficiency, the probability that a rare polynucleotide in the initial genetic material will be represented in the tagged library and, consequently, detected by sequencing is increased. Furthermore, rather than directly address the low conversion efficiency issue of library preparation, most protocols to date call for greater than 1 microgram of DNA as input material. However, when input sample material is limited or detection of polynucleotides with low representation is desired, high conversion efficiency can efficiently sequence the sample and/or to adequately detect such polynucleotides.

[0085] This disclosure provides methods of converting initial polynucleotides into tagged polynucleotides with a conversion efficiency of at least 10%, at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 80% or at least 90%. The methods involve, for example, using any of blunt-end ligation, sticky end ligation, molecular inversion probes, PCR, ligation-based PCR, multiplex PCR, single strand ligation and single strand circularization. The methods can also involve limiting the amount of initial genetic material. For example, the amount of initial genetic material can be less than 1 μ g, less than 100 ng or less than 10 ng. These methods are described in more detail herein.

[0086] Obtaining accurate quantitative and qualitative information about polynucleotides in a tagged library can result in a more sensitive characterization of the initial genetic material. Typically, polynucleotides in a tagged library are amplified and the resulting amplified molecules are sequenced. Depending on the throughput of the sequencing platform used, only a subset of the molecules in the amplified library produce sequence reads. So, for example, the number of amplified molecules sampled for sequencing may be about only 50% of the unique polynucleotides in the tagged library. Furthermore, amplification may be biased in favor of or against certain sequences or certain members of the tagged library. This may distort quantitative measurement of sequences in the tagged library. Also, sequencing platforms can introduce errors in sequencing. For example, sequences can have a per-base error rate of 0.5-1%. Amplification bias and sequencing errors introduce noise into the final sequencing product. This noise can diminish sensitivity of detection. For example, sequence variants whose frequency in the tagged population is less than the sequencing error rate can be mistaken for noise. Also, by providing reads of sequences in greater or less amounts than their actual number in a population, amplification bias can distort measurements of copy number variation.

[0087] This disclosure provides methods of accurately detecting and reading unique polynucleotides in a tagged

pool. In certain embodiments this disclosure provides sequence-tagged polynucleotides that, when amplified and sequenced, provide information that allowed the tracing back, or collapsing, of progeny polynucleotides to the unique tag parent polynucleotide molecule. Collapsing families of amplified progeny polynucleotides reduces amplification bias by providing information about original unique parent molecules. Collapsing also reduces sequencing errors by eliminating from sequencing data mutant sequences of progeny molecules.

[0088] Detecting and reading unique polynucleotides in the tagged library can involve two strategies. In one strategy a sufficiently large subset of the amplified progeny polynucleotide pool is a sequenced such that, for a large percentage of unique tagged parent polynucleotides in the set of tagged parent polynucleotides, there is a sequence read is produced for at least one amplified progeny polynucleotide in a family produced from a unique tagged parent polynucleotide. In a second strategy, the amplified progeny polynucleotide set is sampled for sequencing at a level to produce sequence reads from multiple progeny members of a family derived from a unique parent polynucleotide. Generation of sequence reads from multiple progeny members of a family allows collapsing of sequences into consensus parent sequences.

[0089] So, for example, sampling a number of amplified progeny polynucleotides from the set of amplified progeny polynucleotides that is equal to the number of unique tagged parent polynucleotides in the set of tagged parent polynucleotides (particularly when the number is at least 10,000) will produce, statistically, a sequence read for at least one of progeny of about 68% of the tagged parent polynucleotides in the set, and about 40% of the unique tagged parent polynucleotides in the original set will be represented by at least two progeny sequence reads. In certain embodiments the amplified progeny polynucleotide set is sampled sufficiently so as to produce an average of five to ten sequence reads for each family. Sampling from the amplified progeny set of 10-times as many molecules as the number of unique tagged parent polynucleotides will produce, statistically, sequence information about 99.995% of the families, of which 99.95% of the total families will be covered by a plurality of sequence reads. A consensus sequence can be built from the progeny polynucleotides in each family so as to dramatically reduce the error rate from the nominal per-base sequencing error rate to a rate possibly many orders of magnitude lower. For example, if the sequencer has a random per-base error rate of 1% and the chosen family has 10 reads, a consensus sequence built from these 10 reads would possess an error rate of below 0.0001%. Accordingly, the sampling size of the amplified progeny to be sequenced can be chosen so as to ensure a sequence having a frequency in the sample that is no greater than the nominal per-base sequencing error rate to a rate of the sequencing platform used, has at least 99% chance being represented by at least one read.

[0090] In another embodiment the set of amplified progeny polynucleotides is sampled to a level to produce a high probability e.g., at least 90%, that a sequence represented in the set of tagged parent polynucleotides at a frequency that is about the same as the per base sequencing error rate of the sequencing platform used is covered by at least one sequence read and preferably a plurality of sequence reads. So, for example, if the sequencing platform has a per base error rate of 0.2% in a sequence or set of sequences is represented in the set of tagged parent polynucleotides at a frequency of about 0.2%,

then the number of polynucleotides in the amplified progeny pool that are sequenced can be about X times the number of unique molecules in the set of tagged parent polynucleotides.

[0091] These methods can be combined with any of the noise reduction methods described. Including, for example, qualifying sequence reads for inclusion in the pool of sequences used to generate consensus sequences.

[0092] This information can now be used for both qualitative and quantitative analysis. For example, for quantitative analysis, a measure, e.g., a count, of the amount of tagged parent molecules mapping to a reference sequence is determined. This measure can be compared with a measure of tagged parent molecules mapping to a different genomic region. This comparison can reveal, for example, the relative amounts of parent molecules mapping to each region. This, in turn, provides an indication of copy number variation for molecules mapping to a particular region. For example, if the measure of polynucleotides mapping to a first reference sequence is greater than the measure of polynucleotides mapping to a second reference sequence, this may indicate that the parent population, and by extension the original sample, included polynucleotides from cells exhibiting aneuploidy. The measures can be normalized against a control sample to eliminate various biases.

[0093] For qualitative analysis, sequences from a set of tagged polynucleotides mapping to a reference sequence can be analyzed for variant sequences and their frequency in the population of tagged parent polynucleotides can be measured.

II. Sample Preparation

A. Polynucleotide Isolation and Extraction

[0094] The systems and methods of this disclosure may have a wide variety of uses in the manipulation, preparation, identification and/or quantification of cell free polynucleotides. Examples of polynucleotides include but are not limited to: DNA, RNA, amplicons, cDNA, dsDNA, ssDNA, plasmid DNA, cosmid DNA, high Molecular Weight (MW) DNA, chromosomal DNA, genomic DNA, viral DNA, bacterial DNA, mtDNA (mitochondrial DNA), mRNA, rRNA, tRNA, nRNA, siRNA, snRNA, scaRNA, microRNA, dsRNA, ribozyme, riboswitch and viral RNA (e.g., retroviral RNA).

[0095] Cell free polynucleotides may be derived from a variety of sources including human, mammal, non-human mammal, ape, monkey, chimpanzee, reptilian, amphibian, or avian, sources. Further, samples may be extracted from variety of animal fluids containing cell free sequences, including but not limited to blood, serum, plasma, vitreous, sputum, urine, tears, perspiration, saliva, semen, mucosal excretions, mucus, spinal fluid, amniotic fluid, lymph fluid and the like. Cell free polynucleotides may be fetal in origin (via fluid taken from a pregnant subject), or may be derived from tissue of the subject itself.

[0096] Isolation and extraction of cell free polynucleotides may be performed through collection of bodily fluids using a variety of techniques. In some cases, collection may comprise aspiration of a bodily fluid from a subject using a syringe. In other cases collection may comprise pipetting or direct collection of fluid into a collecting vessel.

[0097] After collection of bodily fluid, cell free polynucleotides may be isolated and extracted using a variety of techniques known in the art. In some cases, cell free DNA may be

isolated, extracted and prepared using commercially available kits such as the Qiagen Qiaamp® Circulating Nucleic Acid Kit protocol. In other examples, Qiagen Qubit™ dsDNA HS Assay kit protocol, Agilent™ DNA 1000 kit, or TruSeq™ Sequencing Library Preparation; Low-Throughput (LT) protocol may be used.

[0098] Generally, cell free polynucleotides are extracted and isolated by from bodily fluids through a partitioning step in which cell free DNAs, as found in solution, are separated from cells and other non soluble components of the bodily fluid. Partitioning may include, but is not limited to, techniques such as centrifugation or filtration. In other cases, cells are not partitioned from cell free DNA first, but rather lysed. In this example, the genomic DNA of intact cells is partitioned through selective precipitation. Cell free polynucleotides, including DNA, may remain soluble and may be separated from insoluble genomic DNA and extracted. Generally, after addition of buffers and other wash steps specific to different kits, DNA may be precipitated using isopropanol precipitation. Further clean up steps may be used such as silica based columns to remove contaminants or salts. General steps may be optimized for specific applications. Non specific bulk carrier polynucleotides, for example, may be added throughout the reaction to optimize certain aspects of the procedure such as yield.

[0099] Isolation and purification of cell free DNA may be accomplished using any means, including, but not limited to, the use of commercial kits and protocols provided by companies such as Sigma Aldrich, Life Technologies, Promega, Affymetrix, IBI or the like. Kits and protocols may also be non-commercially available.

[0100] After isolation, in some cases, the cell free polynucleotides are pre-mixed with one or more additional materials, such as one or more reagents (e.g., ligase, protease, polymerase) prior to sequencing.

[0101] One method of increasing conversion efficiency involves using a ligase engineered for optimal reactivity on single-stranded DNA, such as a ThermoPhage ssDNA ligase derivative. Such ligases bypass traditional steps in library preparation of end-repair and A-tailing that can have poor efficiencies and/or accumulated losses due to intermediate cleanup steps, and allows for twice the probability that either the sense or anti-sense starting polynucleotide will be converted into an appropriately tagged polynucleotide. It also converts double-stranded polynucleotides that may possess overhangs that may not be sufficiently blunt-ended by the typical end-repair reaction. Optimal reactions conditions for this ssDNA reaction are: 1× reaction buffer (50 mM MOPS (pH 7.5), 1 mM DTT, 5 mM MgCl₂, 10 mM KCl). With 50 mM ATP, 25 mg/ml BSA, 2.5 mM MnCl₂, 200 pmol 85 nt ssDNA oligomer and 5 U ssDNA ligase incubated at 65° C. for 1 hour. Subsequent amplification using PCR can further convert the tagged single-stranded library to a double-stranded library and yield an overall conversion efficiency of well above 20%. Other methods of increasing conversion rate, e.g., to above 10%, include, for example, any of the following, alone or in combination: Annealing-optimized molecular-inversion probes, blunt-end ligation with a well-controlled polynucleotide size range, sticky-end ligation or an upfront multiplex amplification step with or without the use of fusion primers.

B. Molecular Bar Coding of Cell Free Polynucleotides

[0102] The systems and methods of this disclosure may also enable the cell free polynucleotides to be tagged or tracked in order to permit subsequent identification and origin of the particular polynucleotide. This feature is in contrast with other methods that use pooled or multiplex reactions and that only provide measurements or analyses as an average of multiple samples. Here, the assignment of an identifier to individual or subgroups of polynucleotides may allow for a unique identity to be assigned to individual sequences or fragments of sequences. This may allow acquisition of data from individual samples and is not limited to averages of samples.

[0103] In some examples, nucleic acids or other molecules derived from a single strand may share a common tag or identifier and therefore may be later identified as being derived from that strand. Similarly, all of the fragments from a single strand of nucleic acid may be tagged with the same identifier or tag, thereby permitting subsequent identification of fragments from the parent strand. In other cases, gene expression products (e.g., mRNA) may be tagged in order to quantify expression, by which the barcode, or the barcode in combination with sequence to which it is attached can be counted. In still other cases, the systems and methods can be used as a PCR amplification control. In such cases, multiple amplification products from a PCR reaction can be tagged with the same tag or identifier. If the products are later sequenced and demonstrate sequence differences, differences among products with the same identifier can then be attributed to PCR error.

[0104] Additionally, individual sequences may be identified based upon characteristics of sequence data for the read themselves. For example, the detection of unique sequence data at the beginning (start) and end (stop) portions of individual sequencing reads may be used, alone or in combination, with the length, or number of base pairs of each sequence read unique sequence to assign unique identities to individual molecules. Fragments from a single strand of nucleic acid, having been assigned a unique identity, may thereby permit subsequent identification of fragments from the parent strand. This can be used in conjunction with bottlenecking the initial starting genetic material to limit diversity.

[0105] Further, using unique sequence data at the beginning (start) and end (stop) portions of individual sequencing reads and sequencing read length may be used, alone or combination, with the use of barcodes. In some cases, the barcodes may be unique as described herein. In other cases, the barcodes themselves may not be unique. In this case, the use of non unique barcodes, in combination with sequence data at the beginning (start) and end (stop) portions of individual sequencing reads and sequencing read length may allow for the assignment of a unique identity to individual sequences. Similarly, fragments from a single strand of nucleic acid having been assigned a unique identity, may thereby permit subsequent identification of fragments from the parent strand.

[0106] Generally, the methods and systems provided herein are useful for preparation of cell free polynucleotide sequences to a down-stream application sequencing reaction. Often, a sequencing method is classic Sanger sequencing. Sequencing methods may include, but are not limited to: high-throughput sequencing, pyrosequencing, sequencing-by-synthesis, single-molecule sequencing, nanopore

sequencing, semiconductor sequencing, sequencing-by-ligation, sequencing-by-hybridization, RNA-Seq (Illumina), Digital Gene Expression (Helicos), Next generation sequencing, Single Molecule Sequencing by Synthesis (SMSS) (Helicos), massively-parallel sequencing, Clonal Single Molecule Array (Solexa), shotgun sequencing, Maxim-Gilbert sequencing, primer walking, and any other sequencing methods known in the art.

C. Assignment of Barcodes to Cell Free Polynucleotide Sequences

[0107] The systems and methods disclosed herein may be used in applications that involve the assignment of unique or non-unique identifiers, or molecular barcodes, to cell free polynucleotides. Often, the identifier is a bar-code oligonucleotide that is used to tag the polynucleotide; but, in some cases, different unique identifiers are used. For example, in some cases, the unique identifier is a hybridization probe. In other cases, the unique identifier is a dye, in which case the attachment may comprise intercalation of the dye into the analyte molecule (such as intercalation into DNA or RNA) or binding to a probe labeled with the dye. In still other cases, the unique identifier may be a nucleic acid oligonucleotide, in which case the attachment to the polynucleotide sequences may comprise a ligation reaction between the oligonucleotide and the sequences or incorporation through PCR. In other cases, the reaction may comprise addition of a metal isotope, either directly to the analyte or by a probe labeled with the isotope. Generally, assignment of unique or non-unique identifiers, or molecular barcodes in reactions of this disclosure may follow methods and systems described by US patent applications 20010053519, 20030152490, 20110160078 and U.S. Pat. No. 6,582,908.

[0108] Often, the method comprises attaching oligonucleotide barcodes to nucleic acid analytes through an enzymatic reaction including but not limited to a ligation reaction. For example, the ligase enzyme may covalently attach a DNA barcode to fragmented DNA (e.g., high molecular-weight DNA). Following the attachment of the barcodes, the molecules may be subjected to a sequencing reaction.

However, other reactions may be used as well. For example, oligonucleotide primers containing barcode sequences may be used in amplification reactions (e.g., PCR, qPCR, reverse-transcriptase PCR, digital PCR, etc.) of the DNA template analytes, thereby producing tagged analytes. After assignment of barcodes to individual cell free polynucleotide sequences, the pool of molecules may be sequenced.

[0109] In some cases, PCR may be used for global amplification of cell free polynucleotide sequences. This may comprise using adapter sequences that may be first ligated to different molecules followed by PCR amplification using universal primers. PCR for sequencing may be performed using any means, including but not limited to use of commercial kits provided by Nugen (WGA kit), Life Technologies, Affymetrix, Promega, Qiagen and the like. In other cases, only certain target molecules within a population of cell free polynucleotide molecules may be amplified. Specific primers, may in conjunction with adapter ligation, may be used to selectively amplify certain targets for downstream sequencing.

[0110] The unique identifiers (e.g., oligonucleotide barcodes, antibodies, probes, etc.) may be introduced to cell free polynucleotide sequences randomly or non-randomly. In some cases, they are introduced at an expected ratio of unique

identifiers to microwells. For example, the unique identifiers may be loaded so that more than about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000, 5000, 10000, 50,000, 100,000, 500,000, 1,000,000, 10,000,000, 50,000,000 or 1,000,000,000 unique identifiers are loaded per genome sample. In some cases, the unique identifiers may be loaded so that less than about 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000, 5000, 10000, 50,000, 100,000, 500,000, 1,000,000, 10,000,000, 50,000,000 or 1,000,000,000 unique identifiers are loaded per genome sample. In some cases, the average number of unique identifiers loaded per sample genome is less than, or greater than, about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000, 5000, 10000, 50,000, 100,000, 500,000, 1,000,000, 10,000,000, 50,000,000 or 1,000,000,000 unique identifiers per genome sample.

[0111] In some cases, the unique identifiers may be a variety of lengths such that each barcode is at least about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000 base pairs. In other cases, the barcodes may comprise less than 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000 base pairs.

[0112] In some cases, unique identifiers may be predetermined or random or semi-random sequence oligonucleotides. In other cases, a plurality of barcodes may be used such that barcodes are not necessarily unique to one another in the plurality. In this example, barcodes may be ligated to individual molecules such that the combination of the bar code and the sequence it may be ligated to creates a unique sequence that may be individually tracked. As described herein, detection of non unique barcodes in combination with sequence data of beginning (start) and end (stop) portions of sequence reads may allow assignment of a unique identity to a particular molecule. The length, or number of base pairs, of an individual sequence read may also be used to assign a unique identity to such a molecule. As described herein, fragments from a single strand of nucleic acid having been assigned a unique identity, may thereby permit subsequent identification of fragments from the parent strand.

[0113] The unique identifiers may be used to tag a wide range of analytes, including but not limited to RNA or DNA molecules. For example, unique identifiers (e.g., barcode oligonucleotides) may be attached to whole strands of nucleic acids or to fragments of nucleic acids (e.g., fragmented genomic DNA, fragmented RNA). The unique identifiers (e.g., oligonucleotides) may also bind to gene expression products, genomic DNA, mitochondrial DNA, RNA, mRNA, and the like.

[0114] In many applications, it may be important to determine whether individual cell free polynucleotide sequences each receive a different unique identifier (e.g., oligonucleotide barcode). If the population of unique identifiers introduced into the systems and methods is not significantly diverse, different analytes may possibly be tagged with identical identifiers. The systems and methods disclosed herein may enable detection of cell free polynucleotide sequences tagged with the same identifier. In some cases, a reference sequences may be included with the population of cell free polynucleotide sequences to be analyzed. The reference sequence may be, for example, a nucleic acid with a known sequence and a known quantity. If the unique identifiers are oligonucleotide barcodes and the analytes are nucleic acids, the tagged analytes may subsequently be sequenced and quantified. These methods may indicate if one or more fragments and/or analytes may have been assigned an identical barcode.

[0115] A method disclosed herein may comprise utilizing reagents necessary for the assignment of barcodes to the analytes. In the case of ligation reactions, reagents including, but not limited to, ligase enzyme, buffer, adapter oligonucleotides, a plurality of unique identifier DNA barcodes and the like may be loaded into the systems and methods. In the case of enrichment, reagents including but not limited to a plurality of PCR primers, oligonucleotides containing unique identifying sequence, or barcode sequence, DNA polymerase, DNTPs, and buffer and the like may be used in preparation for sequencing.

[0116] Generally, the method and system of this disclosure may utilize the methods of U.S. Pat. No. 7,537,897 in using molecular barcodes to count molecules or analytes.

III. Nucleic Acid Sequencing Platforms

[0117] After extraction and isolation of cell free polynucleotides from bodily fluids, cell free sequences may be sequenced. Often, a sequencing method is classic Sanger sequencing. Sequencing methods may include, but are not limited to: high-throughput sequencing, pyrosequencing, sequencing-by-synthesis, single-molecule sequencing, nanopore sequencing, semiconductor sequencing, sequencing-by-ligation, sequencing-by-hybridization, RNA-Seq (Illumina), Digital Gene Expression (Helicos), Next generation sequencing, Single Molecule Sequencing by Synthesis (SMSS) (Helicos), massively-parallel sequencing, Clonal Single Molecule Array (Solexa), shotgun sequencing, Maxim-Gilbert sequencing, primer walking, sequencing using PacBio, SOLiD, Ion Torrent, or Nanopore platforms and any other sequencing methods known in the art.

[0118] In some cases, sequencing reactions various types, as described herein, may comprise a variety of sample processing units. Sample processing units may include but are not limited to multiple lanes, multiple channels, multiple wells, or other mean of processing multiple sample sets substantially simultaneously. Additionally, the sample processing unit may include multiple sample chambers to enable processing of multiple runs simultaneously.

[0119] In some examples, simultaneous sequencing reactions may be performed using multiplex sequencing. In some cases, cell free polynucleotides may be sequenced with at least 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, 100,000 sequencing reactions. In other cases cell free polynucleotides may be sequenced with less than 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, 100,000 sequencing reactions. Sequencing reactions may be performed sequentially or simultaneously. Subsequent data analysis may be performed on all or part of the sequencing reactions. In some cases, data analysis may be performed on at least 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, 100,000 sequencing reactions. In other cases data analysis may be performed on less than 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, 100,000 sequencing reactions.

[0120] In other examples, the number of sequence reactions may provide coverage for a different amounts of the genome. In some cases, sequence coverage of the genome may be at least 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 99%, 99.9% or 100%. In other cases, sequence coverage of the genome may be less than 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 99%, 99.9% or 100%.

[0121] In some examples, sequencing can be performed on cell free polynucleotides that may comprise a variety of different types of nucleic acids. Nucleic acids may be polynucleotides or oligonucleotides. Nucleic acids included, but are not limited to DNA or RNA, single stranded or double stranded or a RNA/cDNA pair.

IV. Polynucleotide Analysis Strategy

[0122] FIG. 8. is a diagram, 800, showing a strategy for analyzing polynucleotides in a sample of initial genetic material. In step 802, a sample containing initial genetic material is provided. The sample can include target nucleic acid in low abundance. For example, nucleic acid from a normal or wild-type genome (e.g., a germline genome) can predominate in a sample that also includes no more than 20%, no more than 10%, no more than 5%, no more than 1%, no more than 0.5% or no more than 0.1% nucleic acid from at least one other genome containing genetic variation, e.g., a cancer genome or a fetal genome, or a genome from another species. The sample can include, for example, cell free nucleic acid or cells comprising nucleic acid. The initial genetic material can constitute no more than 100 ng nucleic acid. This can contribute to proper oversampling of the original polynucleotides by the sequencing or genetic analysis process. Alternatively, the sample can be artificially capped or bottlenecked to reduce the amount of nucleic acid to no more than 100 ng or selectively enriched to analyze only sequences of interest. The sample can be modified to selectively produce sequence reads of molecules mapping to each of one or more selected reference sequences. A sample of 100 ng of nucleic acid can contain about 30,000 human haploid genome equivalents, that is, molecules that, together, provide 30,000-fold coverage of a human genome.

[0123] In step 804 the initial genetic material is converted into a set of tagged parent polynucleotides. Tagging can include attaching sequenced tags to molecules in the initial genetic material. Sequenced tags can be selected so that all unique polynucleotides mapping to the same reference sequence had a unique identifying tag. Conversion can be performed at high efficiency, for example at least 50%.

[0124] In step 806, the set of tagged parent polynucleotides is amplified to produce a set of amplified progeny polynucleotides. Amplification may be, for example, 1,000-fold.

[0125] In step 808, the set of amplified progeny polynucleotides are sampled for sequencing. The sampling rate is chosen so that the sequence reads produced both (1) cover a target number of unique molecules in the set of tagged parent polynucleotides and (2) cover unique molecules in the set of tagged parent polynucleotides at a target coverage fold (e.g., 5- to 10-fold coverage of parent polynucleotides).

[0126] In step 810, the set of sequence reads is collapsed to produce a set of consensus sequences corresponding to unique tagged parent polynucleotides. Sequence reads can be qualified for inclusion in the analysis. For example, sequence reads that fail to meet a quality control scores can be removed from the pool. Sequence reads can be sorted into families representing reads of progeny molecules derived from a particular unique parent molecule. For example, a family of amplified progeny polynucleotides can constitute those amplified molecules derived from a single parent polynucleotide. By comparing sequences of progeny in a family, a consensus sequence of the original parent polynucleotide can be deduced. This produces a set of consensus sequences representing unique parent polynucleotides in the tagged pool.

[0127] In step 812, the set of consensus sequences is analyzed using any of the analytical methods described herein. For example, consensus sequences mapping to a particular reference sequence can be analyzed to detect instances of genetic variation. Consensus sequences mapping to particular reference sequences can be measured and normalized against control samples. Measures of molecules mapping to reference sequences can be compared across a genome to identify areas in the genome in which copy number varies, or heterozygosity is lost.

V. Copy Number Variation Detection

A. Copy Number Variation Detection Using Single Sample

[0128] FIG. 1. is a diagram, 100, showing a strategy for detection of copy number variation in a single subject. As shown herein, copy number variation detection methods can be implemented as follows. After extraction and isolation of cell free polynucleotides in step 102, a single unique sample can be sequenced by a nucleic acid sequencing platform known in the art in step 104. This step generates a plurality of genomic fragment sequence reads. In some cases, these sequences reads may contain barcode information. In other examples, barcodes are not utilized. After sequencing, reads are assigned a quality score. A quality score may be a representation of reads that indicates whether those reads may be useful in subsequent analysis based on a threshold. In some cases, some reads are not of sufficient quality or length to perform the subsequent mapping step. Sequencing reads with a quality score at least 90%, 95%, 99%, 99.9%, 99.99% or 99.999% may be filtered out of the data. In other cases, sequencing reads assigned a quality scored less than 90%, 95%, 99%, 99.9%, 99.99% or 99.999% may be filtered out of the data set. In step 106, the genomic fragment reads that meet a specified quality score threshold are mapped to a reference genome, or a template sequence that is known not to contain copy number variations. After mapping alignment, sequence reads are assigned a mapping score. A mapping score may be a representation or reads mapped back to the reference sequence indicating whether each position is or is not uniquely mappable. In instances, reads may be sequences unrelated to copy number variation analysis. For example, some sequence reads may originate from contaminant polynucleotides. Sequencing reads with a mapping score at least 90%, 95%, 99%, 99.9%, 99.99% or 99.999% may be filtered out of the data set. In other cases, sequencing reads assigned a mapping scored less than 90%, 95%, 99%, 99.9%, 99.99% or 99.999% may be filtered out of the data set.

[0129] After data filtering and mapping, the plurality of sequence reads generates a chromosomal region of coverage. In step 108 these chromosomal regions may be divided into variable length windows or bins. A window or bin may be at least 5 kb, 10, kb, 25 kb, 30 kb, 35, kb, 40 kb, 50 kb, 60 kb, 75 kb, 100 kb, 150 kb, 200 kb, 500 kb, or 1000 kb. A window or bin may also have bases up to 5 kb, 10, kb, 25 kb, 30 kb, 35, kb, 40 kb, 50 kb, 60 kb, 75 kb, 100 kb, 150 kb, 200 kb, 500 kb, or 1000 kb. A window or bin may also be about 5 kb, 10, kb, 25 kb, 30 kb, 35, kb, 40 kb, 50 kb, 60 kb, 75 kb, 100 kb, 150 kb, 200 kb, 500 kb, or 1000 kb.

[0130] For coverage normalization in step 110, each window or bin is selected to contain about the same number of mappable bases. In some cases, each window or bin in a chromosomal region may contain the exact number of map-

pable bases. In other cases, each window or bin may contain a different number of mappable bases. Additionally, each window or bin may be non-overlapping with an adjacent window or bin. In other cases, a window or bin may overlap with another adjacent window or bin. In some cases a window or bin may overlap by at least 1 bp, 2, bp, 3 bp, 4 bp, 5, bp, 10 bp, 20 bp, 25 bp, 50 bp, 100 bp, 200 bp, 250 bp, 500 bp, or 1000 bp. In other cases, a window or bin may overlap by up to 1 by [does this make sense? Less than 1?], 2, bp, 3 bp, 4 bp, 5, bp, 10 bp, 20 bp, 25 bp, 50 bp, 100 bp, 200 bp, 250 bp, 500 bp, or 1000 bp. In some cases a window or bin may overlap by about 1 bp, 2, bp, 3 bp, 4 bp, 5, bp, 10 bp, 20 bp, 25 bp, 50 bp, 100 bp, 200 bp, 250 bp, 500 bp, or 1000 bp.

[0131] In some cases, each of the window regions may be sized so they contain about the same number of uniquely mappable bases. The mappability of each base that comprise a window region is determined and used to generate a mappability file which contains a representation of reads from the references that are mapped back to the reference for each file. The mappability file contains one row per every position, indicating whether each position is or is not uniquely mappable.

[0132] Additionally, predefined windows, known throughout the genome to be hard to sequence, or contain a substantially high GC bias, may be filtered from the data set. For example, regions known to fall near the centromere of chromosomes (i.e., centromeric DNA) are known to contain highly repetitive sequences that may produce false positive results. These regions may be filtered out. Other regions of the genome, such as regions that contain an unusually high concentration of other highly repetitive sequences such as microsatellite DNA, may be filtered from the data set.

[0133] The number of windows analyzed may also vary. In some cases, at least 10, 20, 30, 40, 50, 100, 200, 500, 1000, 2000, 5,000, 10,000, 20,000, 50,000 or 100,000 windows are analyzed. In other cases, the number of widows analyzed is up to 10, 20, 30, 40, 50, 100, 200, 500, 1000, 2000, 5,000, 10,000, 20,000, 50,000 or 100,000 windows are analyzed.

[0134] For an exemplary genome derived from cell free polynucleotide sequences, the next step comprises determining read coverage for each window region. This may be performed using either reads with barcodes, or without barcodes. In cases without barcodes, the pervious mapping steps will provide coverage of different base positions. Sequence reads that have sufficient mapping and quality scores and fall within chromosome windows that are not filtered, may be counted. The number of coverage reads may be assigned a score per each mappable position. In cases involving barcodes, all sequences with the same barcode, physical properties or combination of the two may be collapsed into one read, as they are all derived from the sample parent molecule. This step reduces biases which may have been introduced during any of the preceding steps, such as steps involving amplification. For example, if one molecule is amplified 10 times but another is amplified 1000 times, each molecule is only represented once after collapse thereby negating the effect of uneven amplification. Only reads with unique barcodes may be counted for each mappable position and influence the assigned score.

[0135] Consensus sequences can be generated from families of sequence reads by any method known in the art. Such methods include, for example, linear or non-linear methods of building consensus sequences (such as voting, averaging, statistical, maximum a posteriori or maximum likelihood

detection, dynamic programming, Bayesian, hidden Markov or support vector machine methods, etc.) derived from digital communication theory, information theory, or bioinformatics.

[0136] After the sequence read coverage has been determined, a stochastic modeling algorithm is applied to convert the normalized nucleic acid sequence read coverage for each window region to the discrete copy number states. In some cases, this algorithm may comprise one or more of the following: Hidden Markov Model, dynamic programming, support vector machine, Bayesian network, trellis decoding, Viterbi decoding, expectation maximization, Kalman filtering methodologies and neural networks.

[0137] In step **112**, the discrete copy number states of each window region can be utilized to identify copy number variation in the chromosomal regions. In some cases, all adjacent window regions with the same copy number can be merged into a segment to report the presence or absence of copy number variation state. In some cases, various windows can be filtered before they are merged with other segments.

[0138] In step **114**, the copy number variation may be reported as graph, indicating various positions in the genome and a corresponding increase or decrease or maintenance of copy number variation at each respective position. Additionally, copy number variation may be used to report a percentage score indicating how much disease material (or nucleic acids having a copy number variation) exists in the cell free polynucleotide sample.

B. Copy Number Variation Detection Using Paired Sample

[0139] Paired sample copy number variation detection shares many of the steps and parameters as the single sample approach described herein. However, as depicted in **200** of FIG. 2 of copy number variation detection using paired samples requires comparison of sequence coverage to a control sample rather than comparing it the predicted mappability of the genome. This approach may aid in normalization across windows.

[0140] FIG. 2. is a diagram, **200** showing a strategy for detection of copy number variation in paired subject. As shown herein, copy number variation detection methods can be implemented as follows. In step **204**, a single unique sample can be sequenced by a nucleic acid sequencing platform known in the art after extraction and isolation of the sample in step **202**. This step generates a plurality of genomic fragment sequence reads. Additionally, a sample or control sample is taken from another subject. In some cases, the control subject may be a subject not known to have disease, whereas the other subject may have or be at risk for a particular disease. In some cases, these sequences reads may contain barcode information. In other examples, barcodes are not utilized. After sequencing, reads are assigned a quality score. In some cases, some reads are not of sufficient quality or length to perform the subsequent mapping step. Sequencing reads with a quality score at least 90%, 95%, 99%, 99.9%, 99.99% or 99.999% may be filtered out of the data set. In other cases, sequencing reads assigned a quality scored less than 90%, 95%, 99%, 99.9%, 99.99% or 99.999% may be filtered out of the data set. In step **206**, the genomic fragment reads that meet a specified quality score threshold are mapped to a reference genome, or a template sequence that is known not to contain copy number variations. After mapping alignment, sequence reads are assigned a mapping score. In

instances, reads may be sequences unrelated to copy number variation analysis. For example, some sequence reads may originate from contaminant polynucleotides. Sequencing reads with a mapping score at least 90%, 95%, 99%, 99.9%, 99.99% or 99.999% may be filtered out of the data set. In other cases, sequencing reads assigned a mapping scored less than 90%, 95%, 99%, 99.9%, 99.99% or 99.999% may be filtered out of the data set.

[0141] After data filtering and mapping, the plurality of sequence reads generates a chromosomal region of coverage for each of the test and control subjects. In step **208** these chromosomal regions may be divided into variable length windows or bins. A window or bin may be at least 5 kb, 10, kb, 25 kb, 30 kb, 35, kb, 40 kb, 50 kb, 60 kb, 75 kb, 100 kb, 150 kb, 200 kb, 500 kb, or 1000 kb. A window or bin may also be less than 5 kb, 10, kb, 25 kb, 30 kb, 35, kb, 40 kb, 50 kb, 60 kb, 75 kb, 100 kb, 150 kb, 200 kb, 500 kb, or 1000 kb.

[0142] For coverage normalization in step **210**, each window or bin is selected to contain about the same number of mappable bases for each of the test and control subjects. In some cases, each window or bin in a chromosomal region may contain the exact number of mappable bases. In other cases, each window or bin may contain a different number of mappable bases. Additionally, each window or bin may be non-overlapping with an adjacent window or bin. In other cases, a window or bin may overlap with another adjacent window or bin. In some cases a window or bin may overlap by at least 1 bp, 2, bp, 3 bp, 4 bp, 5, bp, 10 bp, 20 bp, 25 bp, 50 bp, 100 bp, 200 bp, 250 bp, 500 bp, or 1000 bp. In other cases, a window or bin may overlap by less than 1 bp, 2, bp, 3 bp, 4 bp, 5, bp, 10 bp, 20 bp, 25 bp, 50 bp, 100 bp, 200 bp, 250 bp, 500 bp, or 1000 bp.

[0143] In some cases, each of the window regions is sized so they contain about the same number of uniquely mappable bases for each of the test and control subjects. The mappability of each base that comprise a window region is determined and used to generate a mappability file which contains a representation of reads from the references that are mapped back to the reference for each file. The mappability file contains one row per every position, indicating whether each position is or is not uniquely mappable.

[0144] Additionally, predefined windows, known throughout the genome to be hard to sequence, or contain a substantially high GC bias, are filtered from the data set. For example, regions known to fall near the centromere of chromosomes (i.e., centromeric DNA) are known to contain highly repetitive sequences that may produce false positive results. These regions may be filtered. Other regions of the genome, such as regions that contain an unusually high concentration of other highly repetitive sequences such as microsatellite DNA, may be filtered from the data set.

[0145] The number of windows analyzed may also vary. In some cases, at least 10, 20, 30, 40, 50, 100, 200, 500, 1000, 2000, 5,000, 10,000, 20,000, 50,000 or 100,000 windows are analyzed. In other cases, less than 10, 20, 30, 40, 50, 100, 200, 500, 1000, 2000, 5,000, 10,000, 20,000, 50,000 or 100,000 windows are analyzed.

[0146] For an exemplary genome derived from cell free polynucleotide sequences, the next step comprises determining read coverage for each window region for each of the test and control subjects. This may be performed using either reads with barcodes, or without barcodes. In cases without barcodes, the previous mapping steps will provide coverage of different base positions. Sequence reads that have suffi-

cient mapping and quality scores and fall within chromosome windows that are not filtered, may be counted. The number of coverage reads may be assigned a score per each mappable position. In cases involving barcodes, all sequences with the same barcode may be collapsed into one read, as they are all derived from the sample parent molecule. This step reduces biases which may have been introduced during any of the preceding steps, such as steps involving amplification. Only reads with unique barcodes may be counted for each mappable position and influence the assigned score. For this reason, it is important that the barcode ligation step be performed in a manner optimized for producing the lowest amount of bias.

[0147] In determining the nucleic acid read coverage for each window, the coverage of each window can be normalized by the mean coverage of that sample. Using such an approach, it may be desirable to sequence both the test subject and the control under similar conditions. The read coverage for each window may be then expressed as a ratio across similar windows

[0148] Nucleic acid read coverage ratios for each window of the test subject can be determined by dividing the read coverage of each window region of the test sample with read coverage of a corresponding window region of the control sample.

[0149] After the sequence read coverage ratios have been determined, a stochastic modeling algorithm is applied to convert the normalized ratios for each window region into discrete copy number states. In some cases, this algorithm may comprise a Hidden Markov Model. In other cases, the stochastic model may comprise dynamic programming, support vector machine, Bayesian modeling, probabilistic modeling, trellis decoding, Viterbi decoding, expectation maximization, Kalman filtering methodologies, or neural networks.

[0150] In step 212, the discrete copy number states of each window region can be utilized to identify copy number variation in the chromosomal regions. In some cases, all adjacent window regions with the same copy number can be merged into a segment to report the presence or absence of copy number variation state. In some cases, various windows can be filtered before they are merged with other segments.

[0151] In step 214, the copy number variation may be reported as graph, indicating various positions in the genome and a corresponding increase or decrease or maintenance of copy number variation at each respective position. Additionally, copy number variation may be used to report a percentage score indicating how much disease material exists in the cell free poly nucleotide sample.

VI. Rare Mutation Detection

[0152] Rare mutation detection shares similar features as both copy number variation approaches. However, as depicted in FIG. 3, 300, rare mutation detection uses comparison of sequence coverage to a control sample or reference sequence rather than comparing it the relative mappability of the genome. This approach may aid in normalization across windows.

[0153] Generally, rare mutation detection may be performed on selectively enriched regions of the genome or transcriptome purified and isolated in step 302. As described herein, specific regions, which may include but are not limited to genes, oncogenes, tumor suppressor genes, promoters, regulatory sequence elements, non-coding regions, miRNAs,

snRNAs and the like may be selectively amplified from a total population of cell free polynucleotides. This may be performed as herein described. In one example, multiplex sequencing may be used, with or without barcode labels for individual polynucleotide sequences. In other examples, sequencing may be performed using any nucleic acid sequencing platforms known in the art. This step generates a plurality of genomic fragment sequence reads as in step 304. Additionally, a reference sequence is obtained from a control sample, taken from another subject. In some cases, the control subject may be a subject known to not have known genetic aberrations or disease. In some cases, these sequence reads may contain barcode information. In other examples, barcodes are not utilized. After sequencing, reads are assigned a quality score. A quality score may be a representation of reads that indicates whether those reads may be useful in subsequent analysis based on a threshold. In some cases, some reads are not of sufficient quality or length to perform the subsequent mapping step. Sequencing reads with a quality score at least 90%, 95%, 99%, 99.9%, 99.99% or 99.999% may be filtered out of the data set. In other cases, sequencing reads assigned a quality scored at least 90%, 95%, 99%, 99.9%, 99.99% or 99.999% may be filtered out of the data set. In step 306, the genomic fragment reads that meet a specified quality score threshold are mapped to a reference genome, or a reference sequence that is known not to contain rare mutations. After mapping alignment, sequence reads are assigned a mapping score. A mapping score may be a representation or reads mapped back to the reference sequence indicating whether each position is or is not uniquely mappable. In instances, reads may be sequences unrelated to rare mutation analysis. For example, some sequence reads may originate from contaminant polynucleotides. Sequencing reads with a mapping score at least 90%, 95%, 99%, 99.9%, 99.99% or 99.999% may be filtered out of the data set. In other cases, sequencing reads assigned a mapping scored less than 90%, 95%, 99%, 99.9%, 99.99% or 99.999% may be filtered out of the data set.

[0154] For each mappable base, bases that do not meet the minimum threshold for mappability, or low quality bases, may be replaced by the corresponding bases as found in the reference sequence.

[0155] After data filtering and mapping, variant bases found between the sequence reads obtained from the subject and the reference sequence are analyzed.

[0156] For an exemplary genome derived from cell free polynucleotide sequences, the next step comprises determining read coverage for each mappable base position. This may be performed using either reads with barcodes, or without barcodes. In cases without barcodes, the previous mapping steps will provide coverage of different base positions. Sequence reads that have sufficient mapping and quality scores may be counted. The number of coverage reads may be assigned a score per each mappable position. In cases involving barcodes, all sequences with the same barcode may be collapsed into one consensus read, as they are all derived from the sample parent molecule. The sequence for each base is aligned as the most dominant nucleotide read for that specific location. Further, the number of unique molecules can be counted at each position to derive simultaneous quantification at each position. This step reduces biases which may have been introduced during any of the preceding steps, such as

steps involving amplification. Only reads with unique barcodes may be counted for each mappable position and influence the assigned score.

[0157] Once read coverage may be ascertained and variant bases relative to the control sequence in each read are identified, the frequency of variant bases may be calculated as the number of reads containing the variant divided by the total number of reads. This may be expressed as a ratio for each mappable position in the genome.

[0158] For each base position, the frequencies of all four nucleotides, cytosine, guanine, thymine, adenine are analyzed in comparison to the reference sequence. A stochastic or statistical modeling algorithm is applied to convert the normalized ratios for each mappable position to reflect frequency states for each base variant. In some cases, this algorithm may comprise one or more of the following: Hidden Markov Model, dynamic programming, support vector machine, Bayesian or probabilistic modeling, trellis decoding, Viterbi decoding, expectation maximization, Kalman filtering methodologies, and neural networks.

[0159] In step 312, the discrete rare mutation states of each base position can be utilized to identify a base variant with high frequency of variance as compared to the baseline of the reference sequence. In some cases, the baseline might represent a frequency of at least 0.0001%, 0.001%, 0.01%, 0.1%, 1.0%, 2.0%, 3.0%, 4.0% 5.0%, 10%, or 25%. In other cases the baseline might represent a frequency of at least 0.0001%, 0.001%, 0.01%, 0.1%, 1.0%, 2.0%, 3.0%, 4.0% 5.0%, 10%, or 25%. In some cases, all adjacent base positions with the base variant or mutation can be merged into a segment to report the presence or absence of a rare mutation. In some cases, various positions can be filtered before they are merged with other segments.

[0160] After calculation of frequencies of variance for each base position, the variant with largest deviation for a specific position in the sequence derived from the subject as compared to the reference sequence is identified as a rare mutation. In some cases, a rare mutation may be a cancer mutation. In other cases, a rare mutation might be correlated with a disease state.

[0161] A rare mutation or variant may comprise a genetic aberration that includes, but is not limited to a single base substitution, or small indels, transversions, translocations, inversion, deletions, truncations or gene truncations. In some cases, a rare mutation may be at most 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15 or 20 nucleotides in length. On other cases a rare mutation may be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15 or 20 nucleotides in length.

[0162] In step 314, the presence or absence of a mutation may be reflected in graphical form, indicating various positions in the genome and a corresponding increase or decrease or maintenance of a frequency of mutation at each respective position. Additionally, rare mutations may be used to report a percentage score indicating how much disease material exists in the cell free polynucleotide sample. A confidence score may accompany each detected mutation, given known statistics of typical variances at reported positions in non-disease reference sequences. Mutations may also be ranked in order of abundance in the subject or ranked by clinically actionable importance.

VII. Applications

A. Early Detection of Cancer

[0163] Numerous cancers may be detected using the methods and systems described herein. Cancers cells, as most cells, can be characterized by a rate of turnover, in which old cells die and replaced by newer cells. Generally dead cells, in contact with vasculature in a given subject, may release DNA or fragments of DNA into the blood stream. This is also true of cancer cells during various stages of the disease. Cancer cells may also be characterized, dependent on the stage of the disease, by various genetic aberrations such as copy number variation as well as rare mutations. This phenomenon may be used to detect the presence or absence of cancers individuals using the methods and systems described herein.

[0164] For example, blood from subjects at risk for cancer may be drawn and prepared as described herein to generate a population of cell free polynucleotides. In one example, this might be cell free DNA. The systems and methods of the disclosure may be employed to detect rare mutations or copy number variations that may exist in certain cancers present. The method may help detect the presence of cancerous cells in the body, despite the absence of symptoms or other hallmarks of disease.

[0165] The types and number of cancers that may be detected may include but are not limited to blood cancers, brain cancers, lung cancers, skin cancers, nose cancers, throat cancers, liver cancers, bone cancers, lymphomas, pancreatic cancers, skin cancers, bowel cancers, rectal cancers, thyroid cancers, bladder cancers, kidney cancers, mouth cancers, stomach cancers, solid state tumors, heterogeneous tumors, homogenous tumors and the like.

[0166] In the early detection of cancers, any of the systems or methods herein described, including rare mutation detection or copy number variation detection may be utilized to detect cancers. These system and methods may be used to detect any number of genetic aberrations that may cause or result from cancers. These may include but are not limited to 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.

[0167] Additionally, the systems and methods described herein may also be used to help characterize certain cancers. Genetic data produced from the system and methods of this disclosure may allow practitioners to help better characterize a specific form of cancer. Often times, cancers are heterogeneous in both composition and staging. Genetic profile data may allow characterization of specific sub-types of cancer that may be important in the diagnosis or treatment of that specific sub-type. This information may also provide a subject or practitioner clues regarding the prognosis of a specific type of cancer.

B. Cancer Monitoring and Prognosis

[0168] The systems and methods provided herein may be used to monitor already known cancers, or other diseases in a particular subject. This may allow either a subject or practi-

tioner to adapt treatment options in accord with the progress of the disease. In this example, the systems and methods described herein may be used to construct genetic profiles of a particular subject of the course of the disease. In some instances, cancers can progress, becoming more aggressive and genetically unstable. In other examples, cancers may remain benign, inactive or dormant. The system and methods of this disclosure may be useful in determining disease progression.

[0169] Further, the systems and methods described herein may be useful in determining the efficacy of a particular treatment option. In one example, successful treatment options may actually increase the amount of copy number variation or rare mutations detected in subject's blood if the treatment is successful as more cancers may die and shed DNA. In other examples, this may not occur. In another example, perhaps certain treatment options may be correlated with genetic profiles of cancers over time. This correlation may be useful in selecting a therapy. Additionally, if a cancer is observed to be in remission after treatment, the systems and methods described herein may be useful in monitoring residual disease or recurrence of disease.

C. Early Detection and Monitoring of Other Diseases or Disease States

[0170] The methods and systems described herein may not be limited to detection of rare mutations and copy number variations associated with only cancers. Various other diseases and infections may result in other types of conditions that may be suitable for early detection and monitoring. For example, in certain cases, genetic disorders or infectious diseases may cause a certain genetic mosaicism within a subject. This genetic mosaicism may cause copy number variation and rare mutations that could be observed. In another example, the system and methods of the disclosure may also be used to monitor the genomes of immune cells within the body. Immune cells, such as B cells, may undergo rapid clonal expansion upon the presence certain diseases. Clonal expansions may be monitored using copy number variation detection and certain immune states may be monitored. In this example, copy number variation analysis may be performed over time to produce a profile of how a particular disease may be progressing.

[0171] Further, the systems and methods of this disclosure may also be used to monitor systemic infections themselves, as may be caused by a pathogen such as a bacteria or virus. Copy number variation or even rare mutation detection may be used to determine how a population of pathogens are changing during the course of infection. This may be particularly important during chronic infections, such as HIV/AIDs or Hepatitis infections, whereby viruses may change life cycle state and/or mutate into more virulent forms during the course of infection.

[0172] Yet another example that the system and methods of this disclosure may be used for is the monitoring of transplant subjects. Generally, transplanted tissue undergoes a certain degree of rejection by the body upon transplantation. The methods of this disclosure may be used to determine or profile rejection activities of the host body, as immune cells attempt to destroy transplanted tissue. This may be useful in monitoring the status of transplanted tissue as well as altering the course of treatment or prevention of rejection.

[0173] Further, the methods of the disclosure may be used to characterize the heterogeneity of an abnormal condition in

a subject, the method comprising generating a genetic profile of extracellular polynucleotides in the subject, wherein the genetic profile comprises a plurality of data resulting from copy number variation and rare mutation analyses. In some cases, including but not limited to cancer, a disease may be heterogeneous. Disease cells may not be identical. In the example of cancer, some tumors are known to comprise different types of tumor cells, some cells in different stages of the cancer. In other examples, heterogeneity may comprise multiple foci of disease. Again, in the example of cancer, there may be multiple tumor foci, perhaps where one or more foci are the result of metastases that have spread from a primary site.

[0174] The methods of this disclosure may be used to generate or profile, fingerprint or set of data that is a summation of genetic information derived from different cells in a heterogeneous disease. This set of data may comprise copy number variation and rare mutation analyses alone or in combination.

D. Early Detection and Monitoring of Other Diseases or Disease States of Fetal Origin

[0175] Additionally, the systems and methods of the disclosure may be used to diagnose, prognose, monitor or observe cancers or other diseases of fetal origin. That is, these methodologies may be employed in a pregnant subject to diagnose, prognose, monitor or observe cancers or other diseases in a unborn subject whose DNA and other polynucleotides may co-circulate with maternal molecules.

VIII. Terminology

[0176] The terminology used therein is for the purpose of describing particular embodiments only and is not intended to be limiting of a systems and methods of this disclosure. As used herein, the singular forms "a", "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. Furthermore, to the extent that the terms "including", "includes", "having", "has", "with", or variants thereof are used in either the detailed description and/or the claims, such terms are intended to be inclusive in a manner similar to the term "comprising".

[0177] Several aspects of a systems and methods of this disclosure are described above with reference to example applications for illustration. It should be understood that numerous specific details, relationships, and methods are set forth to provide a full understanding of a systems and methods. One having ordinary skill in the relevant art, however, will readily recognize that a systems and methods can be practiced without one or more of the specific details or with other methods. This disclosure is not limited by the illustrated ordering of acts or events, as some acts may occur in different orders and/or concurrently with other acts or events. Furthermore, not all illustrated acts or events are required to implement a methodology in accordance with this disclosure.

[0178] Ranges can be expressed herein as from "about" one particular value, and/or to "about" another particular value. When such a range is expressed, another embodiment includes from the one particular value and/or to the other particular value. Similarly, when values are expressed as approximations, by use of the antecedent "about," it will be understood that the particular value forms another embodiment. It will be further understood that the endpoints of each of the ranges are significant both in relation to the other

endpoint, and independently of the other endpoint. The term "about" as used herein refers to a range that is 15% plus or minus from a stated numerical value within the context of the particular usage. For example, about 10 would include a range from 8.5 to 11.5.

EXAMPLES

Example 1

Prostate Cancer Prognosis and Treatment

[0179] A blood sample is taken from a prostate cancer subject. Previously, an oncologist determines that the subject has stage II prostate cancer and recommends a treatment. Cell free DNA is extracted, isolated, sequenced and analyzed every 6 months after the initial diagnosis.

[0180] Cell free DNA is extracted and isolated from blood using the Qiagen Qubit kit protocol. A carrier DNA is added to increase yields. DNA is amplified using PCR and universal primers. 10 ng of DNA is sequenced using a massively parallel sequencing approach with an Illumina MiSeq personal sequencer. 90% of the subject's genome is covered through sequencing of cell free DNA.

[0181] Sequence data is assembled and analyzed for copy number variation. Sequence reads are mapped and compared to a healthy individual (control). Based on the number of sequence reads, chromosomal regions are divided into 50 kb non overlapping regions. Sequence reads are compared to one another and a ratio is determined for each mappable position.

[0182] A Hidden Markov Model is applied to convert copy numbers into discrete states for each window.

[0183] Reports are generated, mapping genome positions and copy number variation show in FIG. 4A (for a healthy individual) and FIG. 4B for the subject with cancer.

[0184] These reports, in comparison to other profiles of subjects with known outcomes, indicate that this particular cancer is aggressive and resistant to treatment. The cell free tumor burden is 21%. The subject is monitored for 18 months. At month 18, the copy number variation profile begins to increase dramatically, from cell free tumor burden of 21% to 30%. A comparison is done with genetic profiles of other prostate subjects. It is determined that this increase in copy number variation indicates that the prostate cancer is advancing from stage II to stage III. The original treatment regiment as prescribed is no longer treating the cancer. A new treatment is prescribed.

[0185] Further, these reports are submitted and accessed electronically via the internet. Analysis of sequence data occurs at a site other than the location of the subject. The report is generated and transmitted to the subject's location. Via an internet enabled computer, the subject accesses the reports reflecting his tumor burden (FIG. 4C).

Example 2

Prostate Cancer Remission and Recurrence

[0186] A blood sample is taken from a prostate cancer survivor. The subject had previously undergone numerous rounds of chemotherapy and radiation. The subject at the time of testing did not present symptoms or health issues related to the cancer. Standard scans and assays reveal the subject to be cancer free.

[0187] Cell free DNA was extracted and isolated from blood using the Qiagen TruSeq kit protocol. A carrier DNA

was added to increase yields. DNA is amplified using PCR and universal primers. 10 ng of DNA was sequenced using a massively parallel sequencing approach with an Illumina MiSeq personal sequencer. 12mer barcodes were added to individual molecules using a ligation method.

[0188] Sequence data is assembled and analyzed for copy number variation. Sequence reads were mapped and compared to a healthy individual (control). Based on the number of sequence reads, chromosomal regions were divided into 40 kb non overlapping regions. Sequence reads were compared to one another and a ratio is determined for each mappable position.

[0189] Non unique barcoded sequences were collapsed into a single read to help normalize bias from amplification.

[0190] A Hidden Markov Model was applied to convert copy numbers into discrete states for each window.

[0191] Reports were generated, mapping genome positions and copy number variation show in FIG. 5A, for a subject with cancer in remission and FIG. 5B for a subject with cancer in recurrence.

[0192] This reports in comparison to other profiles of subjects with known outcomes indicates that at month 18, rare mutation analysis for copy number variation is detected at cell free tumor burden of 5%. An oncologist prescribes treatment again.

Example 3

Thyroid Cancer and Treatment

[0193] A subject is known to have Stage IV thyroid cancer and undergoes standard treatment, including radiation therapy with I-131. CT scans are inconclusive as to whether the radiation therapy is destroying cancerous masses. Blood is drawn before and after the latest radiation session.

[0194] Cell free DNA is extracted and isolated from blood using the Qiagen Qubit kit protocol. A sample of non specific bulk DNA is added to the sample preparation reactions increase yields.

[0195] It is known that the BRAF gene may be mutated at amino acid position 600 in this thyroid cancer. From population of cell free DNA, BRAF DNA is selectively amplified using primers specific to the gene. 20mer barcodes are added to the parent molecule as a control for counting reads.

[0196] 10 ng of DNA is sequenced using massively parallel sequencing approach with an Illumina MiSeq personal sequencer.

[0197] Sequence data is assembled and analyzed for copy number variation detection. Sequence reads are mapped and compared to a healthy individual (control). Based on the number of sequence reads, as determined by counting the barcode sequences, chromosomal regions are divided into 50 kb non overlapping regions. Sequence reads are compared to one another and a ratio is determined for each mappable position.

[0198] A Hidden Markov Model is applied to convert copy numbers into discrete states for each window.

[0199] A report is generated, mapping genome positions and copy number variation.

[0200] The reports generated before and after treatment are compared. The tumor cell burden percentage jumps from 30% to 60% after the radiation session. The jump in tumor burden is determined to be an increase in necrosis of cancer

tissue versus normal tissue as a result of treatment. Oncologists recommend the subject continue the prescribed treatment.

Example 4

Sensitivity of Rare Mutation Detection

[0201] In order to determine the detection ranges of rare mutation present in a population of DNA, mixing experiments were performed. Sequences of DNA, some containing wildtype copies of the genes TP53, HRAS and MET and some containing copies with rare mutations in the same genes, were mixed together in distinct ratios. DNA mixtures were prepared such that ratios or percentages of mutant DNA to wildtype DNA range from 100% to 0.01%.

[0202] 10 ng of DNA was sequenced for each mixing experiment using a massively parallel sequencing approach with an Illumina MiSeq personal sequencer.

[0203] Sequence data was assembled and analyzed for rare mutation detection. Sequence reads were mapped and compared to a reference sequence (control). Based on the number of sequence reads, the frequency of variance for each mappable position was determined.

[0204] A Hidden Markov Model was applied to convert frequency of variance for each mappable position into discrete states for base position.

[0205] A report was generated, mapping genome base positions and percentage detection of the rare mutation over baseline as determined by the reference sequence (FIG. 6A).

[0206] The results of various mixing experiments ranging from 0.1% to 100% are represented in a logarithmic scale graph, with measured percentage of DNA with a rare mutation graphed as a function of the actual percentage of DNA with a rare mutation (FIG. 6B). The three genes, TP53, HRAS and MET are represented. A strong linear correlation was found between measured and expected rare mutation populations. Additionally, a lower sensitivity threshold of about 0.1% of DNA with a rare mutation in a population of non mutated DNA was found with these experiments (FIG. 6B).

Example 5

Rare Mutation Detection in Prostate Cancer Subject

[0207] A subject was thought to have early stage prostate cancer. Other clinical tests provide inconclusive results. Blood was drawn from the subject and cell free DNA is extracted, isolated, prepared and sequenced.

[0208] A panel of various oncogenes and tumor suppressor genes were selected for selective amplification using a Taq-Man® PCR kit (Invitrogen) using gene specific primers. DNA regions amplified include DNA containing PIK3CA and TP53 genes.

[0209] 10 ng of DNA was sequenced using a massively parallel sequencing approach with an Illumina MiSeq personal sequencer.

[0210] Sequence data was assembled and analyzed for rare mutation detection. Sequence reads are mapped and compared to a reference sequence (control). Based on the number of sequence reads, the frequency of variance for each mappable position was determined.

[0211] A Hidden Markov Model was applied to convert frequency of variance for each mappable position into discrete states for each base position.

[0212] report is generated, mapping genomic base positions and percentage detection of the rare mutation over baseline as determined by the reference sequence (FIG. 7A). Rare mutations are found at an incidence of 5% in two genes, PIK3CA and TP53, respectively, indicating that the subject has an early stage cancer. Treatment is initiated.

[0213] Further, these reports are submitted and accessed electronically via the internet. Analysis of sequence data occurs at a site other than the location of the subject. The report is generated and transmitted to the subject's location. Via an internet enabled computer, the subject accesses the reports reflecting his tumor burden (FIG. 7B).

Example 6

Rare Mutation Detection in Colorectal Cancer Subjects

[0214] A subject is thought to have mid-stage colorectal cancer. Other clinical tests provide inconclusive results. Blood is drawn from the subject and cell free DNA is extracted.

[0215] 10 ng of the cell-free genetic material that is extracted from a single tube of plasma is used. The initial genetic material is converted into a set of tagged parent polynucleotides. The tagging included attaching tags required for sequencing as well as non-unique identifiers for tracking progeny molecules to the parent nucleic acids. The conversion is performed through an optimized ligation reaction as described above and conversion yield is confirmed by looking at the size profile of molecules post-ligation. Conversion yield is measured as the percentage of starting initial molecules that have both ends ligated with tags. Conversion using this approach is performed at high efficiency, for example, at least 50%.

[0216] The tagged library is PCR-amplified and enriched for genes most associated with colorectal cancer, (e.g., KRAS, APC, TP53, etc) and the resulting DNA is sequenced using a massively parallel sequencing approach with an Illumina MiSeq personal sequencer.

[0217] Sequence data is assembled and analyzed for rare mutation detection. Sequence reads are collapsed into familial groups belonging to a parent molecule (as well as error-corrected upon collapse) and mapped using a reference sequence (control). Based on the number of sequence reads, the frequency of rare variations (substitutions, insertions, deletions, etc) and variations in copy number and heterozygosity (when appropriate) for each mappable position is determined.

[0218] report is generated, mapping genomic base positions and percentage detection of the rare mutation over baseline as determined by the reference sequence. Rare mutations are found at an incidence of 0.3-0.4% in two genes, KRAS and FBXW7, respectively, indicating that the subject has residual cancer. Treatment is initiated.

[0219] Further, these reports are submitted and accessed electronically via the internet. Analysis of sequence data occurs at a site other than the location of the subject. The report is generated and transmitted to the subject's location. Via an internet enabled computer, the subject accesses the reports reflecting his tumor burden.

1. 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 generate a plurality of sequencing reads;
- b. filtering out reads that fail to meet a set threshold;
- c. mapping the sequence reads obtained from step (a) to a reference sequence;
- d. quantifying or enumerating mapped reads in two or more predefined regions of the reference sequence;
- 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 other;
 - ii. comparing the normalized numbers obtained in step (i) to normalized numbers obtained from a control sample.

2. A method for detecting a rare mutation in a cell-free or substantially cell free sample obtained from a subject comprising:

- a. sequencing extracellular polynucleotides from a bodily sample from a subject, wherein each of the extracellular polynucleotide generate a plurality of sequencing reads; sequencing extracellular polynucleotides from a bodily sample from a subject, wherein each of the extracellular polynucleotide generate a plurality of sequencing reads;
- b. performing multiplex sequencing on regions or whole-genome sequencing if enrichment is not performed;
- c. filtering out reads that fail to meet a set threshold;
- d. mapping sequence reads derived from the sequencing onto a reference sequence;
- e. identifying a subset of mapped sequence reads that align with a variant of the reference sequence at each mappable base position;
- f. for each mappable base position, calculating a ratio of (a) a number of mapped sequence reads that include a variant as compared to the reference sequence, to (b) a number of total sequence reads for each mappable base position;
- g. normalizing the ratios or frequency of variance for each mappable base position and determining potential rare variant(s) or mutation(s);
- h. and comparing the resulting number for each of the regions with potential rare variant(s) or mutation(s) to similarly derived numbers from a reference sample.

3. A method of characterizing the heterogeneity of an abnormal condition in a subject, the method comprising generating a genetic profile of extracellular polynucleotides in the subject, wherein the genetic profile comprises a plurality of data resulting from copy number variation and rare mutation analyses.

4-49. (canceled)

50. A method comprising: determining copy number variation or rare mutation analysis in a cell-free or substantially cell free sample obtained from a subject using multiplex sequencing.

51-69. (canceled)

70. A system comprising a computer readable medium for performing the following steps: selecting predefined regions in a genome; enumerating number of sequence reads in the predefined regions; normalizing the number of sequence reads across the predefined regions; and determining percent of copy number variation in the predefined regions.

71-76. (canceled)

77. A method comprising:

- a. providing at least one set of tagged parent polynucleotides, and for each set of tagged parent polynucleotides;
- b. amplifying the tagged parent polynucleotides in the set to produce a corresponding set of amplified progeny polynucleotides;
- c. sequencing a subset (including a proper subset) of the set of amplified progeny polynucleotides, to produce a set of sequencing reads; and
- d. collapsing the set of sequencing reads to generate a set of consensus sequences, each consensus sequence corresponding to a unique polynucleotide among the set of tagged parent polynucleotides.

78. The method of claim 77 wherein each polynucleotide in a set is mappable to a reference sequence.

79. The method of claim 77 comprising providing a plurality of sets of tagged parent polynucleotides, wherein each set is mappable to a different reference sequence.

80. The method of claim 77 further comprising: e. analyzing the set of consensus sequences for each set of tagged parent molecules.

81. The method of claim 77 further comprising converting initial starting genetic material into the tagged parent polynucleotides.

82. The method of claim 81 wherein the initial starting genetic material comprises no more than 100 ng of polynucleotides.

83. The method of claim 81 comprising bottlenecking the initial starting genetic material prior to converting.

84. The method of claim 81 comprising converting the initial starting genetic material into tagged parent polynucleotides with a conversion efficiency of at least 10%, at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 80% or at least 90%.

85. The method of claim 81 wherein converting comprises any of blunt-end ligation, sticky end ligation, molecular inversion probes, PCR, ligation-based PCR, single strand ligation and single strand circularization.

86. The method of claim 81 wherein the initial starting genetic material is cell-free nucleic acid.

87. The method of claim 86 wherein a plurality of the reference sequences are from the same genome.

88. The method of claim 77 wherein each tagged parent polynucleotide in the set is uniquely tagged.

89-103. (canceled)

104. A method comprising detecting genetic variation in initial starting genetic material with a sensitivity of at least 5%, at least 1%, at least 0.5%, at least 0.1% or at least 0.05%.

105. (canceled)

106. A system comprising a computer readable medium for performing the following steps:

- a. providing at least one set of tagged parent polynucleotides, and for each set of tagged parent polynucleotides;
- b. amplifying the tagged parent polynucleotides in the set to produce a corresponding set of amplified progeny polynucleotides;
- c. sequencing a subset (including a proper subset) of the set of amplified progeny polynucleotides, to produce a set of sequencing reads;
- d. collapsing the set of sequencing reads to generate a set of consensus sequences, each consensus sequence corresponding to a unique polynucleotide among the set of tagged parent polynucleotides and, optionally,

e. analyzing the set of consensus sequences for each set of tagged parent molecules.

107. A method comprising detecting the presence or absence of genetic alteration and/or amount of genetic variation in an individual, wherein the detecting is performed with the aid of sequencing of cell-free nucleic acid, wherein at least 10% of the individual's genome is sequenced.

108-133. (canceled)

* * * * *