METHODS AND SYSTEMS FOR DETECTING GENETIC VARIANTS

CROSS-REFERENCE

[0001] This application claims the benefit under 35 U.S.C. § 119(e) of U.S. Provisional Application No. 61/921,456, filed December 28, 2013, and U.S. Provisional Application No. 61/948,509, filed March 5, 2014, each of which is entirely incorporated herein by reference.

BACKGROUND

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

[0003] Early detection and monitoring of genetic diseases, such as cancer, is often useful and needed in the successful treatment or management of the disease. One approach may include the monitoring of a sample derived from cell-free nucleic acids, a population of polynucleotides that can be found in different types of bodily fluids. In some cases, disease may be characterized or detected based on detection of genetic aberrations, such as copy number variation and/or sequence variation of one or more nucleic acid sequences, or the development of other certain rare genetic alterations. Cell-free DNA (cfDNA) 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.

[0004] In particular, many methods have been developed for accurate copy number variation estimation, especially for heterogeneous genomic samples, such as tumor-derived gDNA or for cfDNA for many applications (*e.g.*, prenatal, transplant, immune, metagenomics or cancer diagnostics). Most of these methods include sample preparation whereby the original nucleic acids are converted into a sequenceable library, followed by massively parallel sequencing, and finally bioinformatics to estimate copy number variation at one or more loci.

SUMMARY

[0005] Although many of these methods are able to reduce or combat the errors introduced by the sample preparation and sequencing processes for all molecules that are converted and sequenced, these methods are not able to infer the counts of molecules that were converted but not sequenced. Since this count of converted by unsequenced molecules can be highly variable from genomic region to region, these counts can dramatically and adversely affect the sensitivity that can be achieved.

[0006] To address this issue, input double-stranded deoxyribonucleic acid (DNA) can be converted by a process that tags both halves of the individual double-stranded molecule, in some cases differently. This can be performed using a variety of techniques, including ligation of hairpin, bubble, or forked adapters or other adaptors having double-stranded and single stranded segments (the unhybridized portion of a bubble, forked or hairpin adapter are deemed single-stranded herein). If tagged correctly, each original Watson and Crick (i.e., strand) side of the input double-stranded DNA molecule can be differently tagged and identified by the sequencer and subsequent bioinformatics. For all molecules in a particular region, counts of molecules where both Watson and Crick sides were recovered ("Pairs") versus those where only one half was recovered ("Singlets") can be recorded. The number of unseen molecules can be estimated based on the number of Pairs and Singlets detected.

[0007] An aspect of the present disclosure provides a method for detecting and/or quantifying rare deoxyribonucleic acid (DNA) in a heterogeneous population of original DNA fragments, comprising tagging the original DNA fragments in a single reaction using a library of a plurality of different tags such that greater than 30% of the fragments are tagged at both ends, wherein each of the tags comprises a molecular barcode. The single reaction can be in a single reaction vessel. Greater than 50% of the fragments can be tagged at both ends. The plurality of different tags can be no more than any of 100, 500, 1000, 10,000 or 100,000 different tags.

[0008] Another aspect provides a set of library adaptors that can be used to tag the molecules of interest (e.g., by ligation, hybridization, etc.). The set of library adaptors can comprise plurality of polynucleotide molecules with molecular barcodes, wherein the plurality of polynucleotide molecules are less than or equal to 80 nucleotide bases in length, wherein the molecular barcodes are at least 4 nucleotide bases in length, and wherein (a) the molecular barcodes are different from one another and have an edit distance of at least 1 between one another; (b) the molecular barcodes are located at least one nucleotide base

away from a terminal end of their respective polynucleotide molecules; (c) optionally, at least one terminal base is identical in all of the polynucleotide molecules; and (d) none of the polynucleotide molecules contains a complete sequencer motif.

[0009] In some embodiments, the library adaptors (or adapters) are identical to one another but for the molecular barcodes. In some embodiments, each of the plurality of library adaptors comprises at least one double-stranded portion and at least one single-stranded portion (e.g., a non-complementary portion or an overhang). In some embodiments, the double-stranded portion has a molecular barcode selected from a collection of different molecular barcodes. In some embodiments, the given molecular barcode is a randomer. In some embodiments, each of the library adaptors further comprises a strand-identification barcode on the at least one single-stranded portion. In some embodiments, the strand-identification barcode includes at least 4 nucleotide bases. In some embodiments, the single-stranded portion has a partial sequencer motif. In some embodiments, the library adaptors do not include a complete sequencer motif.

[0010] In some embodiments, none of the library adaptors contains a sequence for hybridizing to a flow cell or forming a hairpin for sequencing.

[0011] In some embodiments, all of the library adaptors have a terminal end with nucleotide(s) that are the same. In some embodiments, the identical terminal nucleotide(s) are over two or more nucleotide bases in length.

[0012] In some embodiments, each of the library adapters is Y-shaped, bubble shaped or hairpin shaped. In some embodiments, none of the library adapters contains a sample identification motif. In some embodiments, each of the library adapters comprises a sequence that is selectively hybridizable to a universal primer. In some embodiments, each of the library adapters comprises a molecular barcode that is at least 5, 6, 7, 8, 9 and 10 nucleotide bases in length. In some embodiments, each of the library adapters is from 10 nucleotide bases to 80 in length, or 30 to 70 nucleotide bases in length, or 40 to 60 nucleotide bases in length. In some embodiments, at least 1, 2, 3, or 4 terminal bases are identical in all of the library adaptors. In some embodiments, at least 4 terminal bases are identical in all of the library adaptors.

[0013] In some embodiments, the edit distance of the molecular barcodes of the library adapters is a Hamming distance. In some embodiments, the edit distance is at least 1, 2, 3, 4 or 5. In some embodiments, the edit distance is with respect to individual bases of the plurality of polynucleotide molecules. In some embodiments, the molecular barcodes are

located at least 10 nucleotide base away from a terminal end of an adapter. In some embodiments, the plurality of library adapters includes at least 2, 4, 6, 8, 10, 20, 30, 40 or 50 different molecular barcodes, or from 2-100, 4-80, 6-60 or 8-40 different molecular barcodes. In any of the embodiments herein, there are more polynucleotides (e.g., cfDNA fragments) to be tagged than there are different molecular barcodes such that the tagging is not unique.

[0014] In some embodiments, the terminal end of an adaptor is configured for ligation (e.g., to a target nucleic acid molecule). In some embodiments, the terminal end of an adaptor is a blunt end.

[0015] In some embodiments, the adaptors are purified and isolated. In some embodiments, the library comprises one or more non-naturally occurring bases.

[0016] In some embodiments, the polynucleotide molecules comprise a primer sequence positioned 5' with respect to the molecular barcodes.

[0017] In some embodiments, the set of library adaptors consists essentially of the plurality of polynucleotide molecules.

[0018] In another aspect, a method comprises (a) tagging a collection of polynucleotides with a plurality of polynucleotide molecules from a library of adaptors to create a collection of tagged polynucleotides; and (b) amplifying the collection of tagged polynucleotides in the presence of sequencing adaptors, wherein the sequencing adaptors have primers with nucleotide sequences that are selectively hybridizable to complementary sequences in the plurality of polynucleotide molecules. The library of adaptors may be as described above or elsewhere herein. In some embodiments, each of the sequencer adaptors further comprises an index tag, which can be a sample identification motif.

[0019] Another aspect, provides a method for detecting and/or quantifying rare DNA in a heterogeneous population of original DNA fragments, wherein the rare DNA has a concentration that is less than 1%, the method comprising (a) tagging the original DNA fragments in a single reaction such that greater than 30% of the original DNA fragments are tagged at both ends with library adaptors that comprise molecular barcodes, thereby providing tagged DNA fragments; (b) performing high-fidelity amplification on the tagged DNA fragments; (c) optionally, selectively enriching a subset of the tagged DNA fragments; (d) sequencing one or both strands of the tagged, amplified and optionally selectively enriched DNA fragments to obtain sequence reads comprising nucleotide sequences of the molecular barcodes and at least a portion of the original DNA fragments; (e) from the sequence reads, determining consensus reads that are representative of single-strands of the

original DNA fragments; and (f) quantifying the consensus reads to detect and/or quantify the rare DNA at a specificity that is greater than 99.9%.

[0020] In some embodiments, (e) comprises comparing sequence reads having the same or similar molecular barcodes and the same or similar end of fragment sequences. In some embodiments, the comparing further comprises performing a phylogentic analysis on the sequence reads having the same or similar molecular barcodes. In some embodiments, the molecular barcodes include a barcode having an edit distance of up to 3. In some embodiments, the end of fragment sequence includes fragment sequences having an edit distance of up to 3.

[0021] In some embodiments, the method further comprises sorting sequence reads into paired reads and unpaired reads, and quantifying a number of paired reads and unpaired reads that map to each of one or more genetic loci.

[0022] In some embodiments, the tagging occurs by having an excess amount of library adaptors as compared to original DNA fragments. In some embodiments, n the excess is at least a 5-fold excess. In some embodiments, the tagging comprises using a ligase. In some embodiments, the tagging comprises attachment to blunt ends.

[0023] In some embodiments, the method further comprises binning the sequence reads according to the molecular barcodes and sequence information from at least one end of each of the original DNA fragments to create bins of single stranded reads. In some embodiments, the method further comprises, in each bin, determining a sequence of a given original DNA fragment among the original DNA fragments by analyzing sequence reads. In some embodiments, the method further comprises detecting and/or quantifying the rare DNA by comparing a number of times each base occurs at each position of a genome represented by the tagged, amplified, and optionally enriched DNA fragments.

[0024] In some embodiments, the library adaptors do not contain complete sequencer motifs. In some embodiments, the method further comprises selectively enriching a subset of the tagged DNA fragments. In some embodiments, the method further comprises, after enriching, amplifying the enriched tagged DNA fragments in the presence of sequencing adaptors comprising primers. In some embodiments, (a) provides tagged DNA fragments having from 2 to 1000 different combinations of molecular barcodes.

[0025] In some embodiments, the DNA fragments are tagged with polynucleotide molecules from a library of adaptors as described above or elsewhere herein.

[0026] In another aspect, a method for processing and/or analyzing a nucleic acid sample of a subject comprises (a) exposing polynucleotide fragments from the nucleic acid sample to a set of library adaptors to generate tagged polynucleotide fragments; and (b) subjecting the tagged polynucleotide fragments to nucleic acid amplification reactions under conditions that yield amplified polynucleotide fragments as amplification products of the tagged polynucleotide fragments. The set of library adaptors comprises a plurality of polynucleotide molecules with molecular barcodes, wherein the plurality of polynucleotide molecules are less than or equal to 80 nucleotide bases in length, wherein the molecular barcodes are at least 4 nucleotide bases in length, and wherein (1) the molecular barcodes are different from one another and have an edit distance of at least 1 between one another; (2) the molecular barcodes are located at least one nucleotide base away from a terminal end of their respective polynucleotide molecules; (3) optionally, at least one terminal base is identical in all of the polynucleotide molecules; and (4) none of the polynucleotide molecules contains a complete sequencer motif.

[0027] In some embodiments, the method further comprises determining nucleotide sequences of the amplified tagged polynucleotide fragments. In some embodiments, the nucleotide sequences of the amplified tagged polynucleotide fragments are determined without polymerase chain reaction (PCR). In some embodiments, the method further comprises analyzing the nucleotide sequences with a programmed computer processor to identify one or more genetic variants in the nucleotide sample of the subject. In some embodiments, the one or more genetic variants are selected from the group consisting of base change(s), insertion(s), repeat(s), deletion(s), copy number variation(s) and transversion(s). In some embodiments, the one or more genetic variants include one or more tumor associated genetic alterations.

[0028] In some embodiments, the subject has or is suspected of having a disease. In some embodiments, the disease is cancer. In some embodiments, the method further comprises collecting the nucleic acid sample from the subject. In some embodiments, the nucleic acid sample is collected from a location selected from the group consisting of blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool, cerebral spinal fluid and tears of the subject. In some embodiments, the nucleic acid sample is a cell-free nucleic acid sample. In some embodiments, the nucleic acid sample is collected from no more than 100 nanograms (ng) of double-stranded polynucleotide molecules of the subject.

[0029] In some embodiments, the polynucleotide fragments comprise double-stranded polynucleotide molecules. In some embodiments, in (a), the plurality of polynucleotide molecules couple to the polynucleotide fragments via blunt end ligation, sticky end ligation, molecular inversion probes, PCR, ligation-based PCR, multiplex PCR, single stranded ligation, and single stranded circularization. In some embodiments, exposing the polynucleotide fragments of the nucleic acid sample to the plurality of polynucleotide molecules yields the tagged polynucleotide fragments with a conversion efficiency of at least 10%. In some embodiments, any of at least 5%, 6%, 7%, 8%, 9%, 10%, 20%, or 25% of the tagged polynucleotide fragments share a common polynucleotide molecule or sequence. In some embodiments, the method further comprises generating the polynucleotide fragments from the nucleic acid sample.

[0030] In some embodiments, the subjecting comprises amplifying the tagged polynucleotide fragments from sequences corresponding to genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

[0031] In another aspect, a method comprises (a) generating a plurality of sequence reads from a plurality of polynucleotide molecules, wherein the plurality of polynucleotide molecules cover genomic loci of a target genome, wherein the genomic loci correspond to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1; (b) grouping with a computer processor the plurality of sequence reads into families, wherein each family comprises sequence reads

from one of the template polynucleotides; (c) for each of the families, merging sequence reads to generate a consensus sequence; (d) calling the consensus sequence at a given genomic locus among the genomic loci; and (e) detecting at the given genomic locus any of genetic variants among the calls, frequency of a genetic alteration among the calls, total number of calls, and total number of alterations among the calls.

[0032] In some embodiments, each family comprises sequence reads from only one of the template polynucleotides. In some embodiments, the given genomic locus comprises at least one nucleic acid base. In some embodiments, the given genomic locus comprises a plurality of nucleic acid bases. In some embodiments, the calling comprises calling at least one nucleic acid base at the given genomic locus. In some embodiments, the calling comprises calling a plurality of nucleic acid bases at the given genomic locus. In some embodiments, the calling comprises any one of phylogenetic analysis, voting, weighing, assigning a probability to each read at the locus in a family and calling the base with the highest probability.

[0033] In some embodiments, the method further comprises performing (d)-(e) at an additional genomic locus among the genomic loci. In some embodiments, the method further comprises determining a variation in copy number at one of the given genomic locus and additional genomic locus based on counts at the given genomic locus and additional genomic locus.

[0034] In some embodiments, the grouping comprises classifying the plurality of sequence reads into families by identifying (i) different molecular barcodes coupled to the plurality of polynucleotide molecules and (ii) similarities between the plurality of sequence reads, wherein each family includes a plurality of nucleic acid sequences that are associated with a different combination of molecular barcodes and similar or identical sequence reads. Different molecular barcodes have different sequences.

[0035] In some embodiments, the consensus sequence is generated by evaluating a quantitative measure or a statistical significance level for each of the sequence reads. In some embodiments, the quantitative measure comprises use of a binomial distribution, exponential distribution, beta distribution, or empirical distribution. In some embodiments, the method further comprises mapping the consensus sequence to the target genome. In some embodiments, the plurality of genes includes at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50 or all of the plurality of genes selected from the group.

[0036] Another aspect of the present disclosure provides a method, comprising (a) providing template polynucleotide molecules and a set of library adaptors in a single reaction vessel, wherein the library adaptors are polynucleotide molecules that have different molecular barcodes (e.g., from 2 to 1,000 different molecular barcodes), and wherein none of the library adaptors contains a complete sequencer motif; (b) in the single reaction vessel, coupling the library adaptors to the template polynucleotide molecules at an efficiency of at least 10%, thereby tagging each template polynucleotide with a tagging combination that is among a plurality of different tagging combinations (e.g., 4 to 1,000,000 different tagging combinations), to produce tagged polynucleotide molecules; (c) subjecting the tagged polynucleotide molecules to an amplification reaction under conditions that yield amplified polynucleotide molecules as amplification products of the tagged polynucleotide molecules; and (d) sequencing the amplified polynucleotide molecules.

[0037] In some embodiments, the template polynucleotide molecules are blunt ended or sticky-ended. In some embodiments, the library adaptors are identical but for the molecular barcodes. In some embodiments, each of the library adaptors has a double stranded portion and at least one single-stranded portion. In some embodiments, the double-stranded portion has a molecular barcode among the molecular barcodes. In some embodiments, each of the library adaptors further comprises a strand-identification barcode on the at least one single-stranded portion. In some embodiments, the single-stranded portion has a partial sequencer motif. In some embodiments, the library adaptors have a sequence of terminal nucleotides that are the same. In some embodiments, the template polynucleotide molecules are double-stranded. In some embodiments, the library adaptors couple to both ends of the template polynucleotide molecules.

[0038] In some embodiments, subjecting the tagged polynucleotide molecules to the amplification reaction comprises non-specifically amplifying the tagged polynucleotide molecules.

[0039] In some embodiments, the amplification reaction comprises use of a priming site to amplify each of the tagged polynucleotide molecules. In some embodiments, the priming site is a primer. In some embodiments, the primer is a universal primer. In some embodiments, the priming site is a nick.

[0040] In some embodiments, the method further comprises, prior to (e), (i) separating polynucleotide molecules comprising one or more given sequences from the amplified

polynucleotide molecules, to produce enriched polynucleotide molecules; and (ii) amplifying the enriched polynucleotide molecules with sequencing adaptors.

[0041] In some embodiments, the efficiency is at least 30%, 40%, or 50%. In some embodiments, the method further comprises identifying genetic variants upon sequencing the amplified polynucleotide molecules. In some embodiments, the sequencing comprises (i) subjecting the amplified polynucleotide molecules to an additional amplification reaction under conditions that yield additional amplified polynucleotide molecules as amplification products of the amplified polynucleotide molecules, and (ii) sequencing the additional amplified polynucleotide molecules. In some embodiments, the additional amplification is performed in the presence of sequencing adaptors.

[0042] In some embodiments, (b) and (c) are performed without aliquoting the tagged polynucleotide molecules. In some embodiments, the tagging is non-unique tagging.

[0043] Another aspect, provides a system for analyzing a target nucleic acid molecule of a subject, comprising a communication interface that receives nucleic acid sequence reads for a plurality of polynucleotide molecules that cover genomic loci of a target genome; computer memory that stores the nucleic acid sequence reads for the plurality of polynucleotide molecules received by the communication interface; and a computer processor operatively coupled to the communication interface and the memory and programmed to (i) group the plurality of sequence reads into families, wherein each family comprises sequence reads from one of the template polynucleotides, (ii) for each of the families, merge sequence reads to generate a consensus sequence, (iii) call the consensus sequence at a given genomic locus among the genomic loci, and (iv) detect at the given genomic locus any of genetic variants among the calls, frequency of a genetic alteration among the calls, total number of calls; and total number of alterations among the calls, wherein the genomic loci correspond to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

[0044] In another aspect, a set of oligonucleotide molecules that selectively hybridize to at least 5 genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

[0045] In some embodiments, the oligonucleotide molecules are from 10-200 bases in length. In some embodiments, the oligonucleotide molecules selectively hybridize to exon regions of the at least 5 genes. In some embodiments, the oligonucleotide molecules selectively hybridize to at least 30 exons in the at least 5 genes. In some embodiments, multiple oligonucleotide molecules selectively hybridize to each of the at least 30 exons. In some embodiments, the oligonucleotide molecules that hybridize to each exon have sequences that overlap with at least 1 other oligonucleotide molecule.

[0046] In another aspect, a kit comprises a first container containing a plurality of library adaptors each having a different molecular barcode; and a second container containing a plurality of sequencing adaptors, each sequencing adaptor comprising at least a portion of a sequencer motif and optionally a sample barcode. The library adaptors can be as described above or elsewhere herein.

[0047] In some embodiments, the sequencing adaptor comprises the sample barcode. In some embodiments, the library adaptors are blunt ended and Y-shaped, and are less than or equal to 80 nucleic acid bases in length. In some embodiments, the sequencing adaptor is up to 70 bases from end to end.

[0048] In another aspect, a method for detecting sequence variants in a cell free DNA sample, comprising detecting rare DNA at a concentration less than 1% with a specificity that is greater than 99.9%.

[0049] In another aspect, a method comprises detecting genetic variants in a sample comprising DNA with a detection limit of at least 1% and specificity greater than 99.9%. In some embodiments, the method further comprises converting cDNA (e.g. cfDNA) into adaptor tagged DNA with a conversion efficiency of at least 30%, 40%, or 50% and reducing sequencing noise (or distortion) by eliminating false positive sequence reads.

[0050] Another aspect provides a method, comprising (a) providing a sample comprising a set of double-stranded polynucleotide molecules, each double-stranded polynucleotide molecule including first and second complementary strands; (b) tagging the double-stranded polynucleotide molecules with a set of duplex tags, wherein each duplex tag differently tags the first and second complementary strands of a double-stranded polynucleotide molecule in the set; (c) sequencing at least some of the tagged strands to produce a set of sequence reads; (d) reducing and/or tracking redundancy in the set of sequence reads; (e) sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads; (f) determining quantitative measures of (i) the paired reads and (ii) the unpaired reads that map to each of one or more genetic loci; and (g) estimating with a programmed computer processor a quantitative measure of total double-stranded polynucleotide molecules in the set that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads mapping to each locus.

[0051] In some embodiments, the method further comprises (h) detecting copy number variation in the sample by determining a normalized total quantitative measure determined in step (g) at each of the one or more genetic loci and determining copy number variation based on the normalized measure. In some embodiments, the sample comprises double-stranded polynucleotide molecules sourced substantially from cell-free nucleic acids. In some embodiments, the duplex tags are not sequencing adaptors.

[0052] In some embodiments, reducing redundancy in the set of sequence reads comprises collapsing sequence reads produced from amplified products of an original polynucleotide molecule in the sample back to the original polynucleotide molecule. In some embodiments, the method further comprises determining a consensus sequence for the original polynucleotide molecule. In some embodiments, the method further comprises identifying polynucleotide molecules at one or more genetic loci comprising a sequence variant. In some embodiments, the method further comprises determining a quantitative measure of paired reads that map to a locus, wherein both strands of the pair comprise a sequence variant. In some embodiments, the method further comprises determining a

quantitative measure of paired molecules in which only one member of the pair bears a sequence variant and/or determining a quantitative measure of unpaired molecules bearing a sequence variant. In some embodiments, the sequence variant is selected from the group consisting of a single nucleotide variant, an indel, a transversion, a translocation, an inversion, a deletion, a chromosomal structure alteration, a gene fusion, a chromosome fusion, a gene truncation, a gene amplification, a gene duplication and a chromosomal lesion.

[0053] Another aspect provides a system comprising a computer readable medium comprising machine-executable code that, upon execution by a computer processor, implements a method comprising (a) receiving into memory a set of sequence reads of polynucleotides tagged with duplex tags; (b) reducing and/or tracking redundancy in the set of sequence reads; (c) sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads; (d) determining quantitative measures of (i) the paired reads and (ii) the unpaired reads that map to each of one or more genetic loci; and (e) estimating a quantitative measure of total double-stranded polynucleotide molecules in the set that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads mapping to each locus.

[0054] Another aspect provides a method, comprising (a) providing a sample comprising a set of double-stranded polynucleotide molecules, each double-stranded polynucleotide molecule including first and second complementary strands; (b) tagging the double-stranded polynucleotide molecules with a set of duplex tags, wherein each duplex tag differently tags the first and second complementary strands of a double-stranded polynucleotide molecule in the set; (c) sequencing at least some of the tagged strands to produce a set of sequence reads; (d) reducing and/or tracking redundancy in the set of sequence reads; (e) sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented

among the sequence reads in the set of sequence reads; and (f) determining quantitative measures of at least two of (i) the paired reads, (ii) the unpaired reads that map to each of one or more genetic loci, (iii) read depth of the paired reads and (iv) read depth of unpaired reads.

[0055] In some embodiments, (f) comprises determining quantitative measures of at least three of (i)-(iv). In some embodiments, (f) comprises determining quantitative measures of all of (i)-(iv). In some embodiments, the method further comprises (g) estimating with a programmed computer processor a quantitative measure of total double-stranded polynucleotide molecules in the set that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads and their read depths mapping to each locus.

[0056] In another aspect, a method comprises (a) tagging control parent polynucleotides with a first tag set to produce tagged control parent polynucleotides, wherein the first tag set comprises a plurality of tags, wherein each tag in the first tag set comprises a same control tag and an identifying tag, and wherein the tag set comprises a plurality of different identifying tags; (b) tagging test parent polynucleotides with a second tag set to produce tagged test parent polynucleotides, wherein the second tag set comprises a plurality of tags, wherein each tag in the second tag set comprises a same test tag that is distinguishable from the control tag and an identifying tag, and wherein the second tag set comprises a plurality of different identifying tags; (c) mixing tagged control parent polynucleotides with tagged test parent polynucleotides to form a pool; (d) amplifying tagged parent polynucleotides in the pool to form a pool of amplified, tagged polynucleotides; (e) sequencing amplified, tagged polynucleotides in the amplified pool to produce a plurality of sequence reads; (f) grouping sequence reads into families, each family comprising sequence reads generated from a same parent polynucleotide, which grouping is optionally based on information from an identifying tag and from start/end sequences of the parent polynucleotides, and, optionally, determining a consensus sequence for each of a plurality of parent polynucleotides from the plurality of sequence reads in a group; (g) classifying each family or consensus sequence as a control parent polynucleotide or as a test parent polynucleotide based on having a test tag or a control tag; (h) determining a quantitative measure of control parent polynucleotides and control test polynucleotides mapping to each of at least two genetic loci; and (i) determining copy number variation in the test parent polynucleotides at at least one locus based on relative quantity of test parent polynucleotides and control parent polynucleotides mapping to the at least one locus.

[0057] In another aspect, a method comprises (a) generating a plurality of sequence reads from a plurality of template polynucleotides, each polynucleotide mapped to a genomic locus; (b) grouping the sequence reads into families, each family comprising sequence reads generated from one of the template polynucleotides; (c) calling a base (or sequence) at the genomic locus for each of the families; (d) detecting at the genomic locus any of genomic alterations among the calls, frequency of a genetic alteration among the calls, total number of calls and total number of alterations among the calls.

[0058] In some embodiments, calling comprises any of phylogenetic analysis, voting, weighing, assigning a probability to each read at the locus in a family, and calling the base with the highest probability. In some embodiments, the method is performed at two loci, comprising determining CNV at one of the loci based on counts at each of the loci.

[0059] Another aspect provides a method for determining a quantitative measure indicative of a number of individual double-stranded DNA fragments in a sample comprising (a) determining a quantitative measure of individual DNA molecules for which both strands are detected; (b) determining a quantitative measure of individual DNA molecules for which only one of the DNA strands are detected; (c) inferring from (a) and (b) above a quantitative measure of individual DNA molecules for which neither strand was detected; and (d) using (a)-(c) determining the quantitative measure indicative of a number of individual double-stranded DNA fragments in the sample.

[0060] In some embodiments, the method further comprises detecting copy number variation in the sample by determining a normalized quantitative measure determined in step (d) at each of one or more genetic loci and determining copy number variation based on the normalized measure. In some embodiments, the sample comprises double-stranded polynucleotide molecules sourced substantially from cell-free nucleic acids.

[0061] In some embodiments, determining the quantitative measure of individual DNA molecules comprises tagging the DNA molecules with a set of duplex tags, wherein each duplex tag differently tags complementary strands of a double-stranded DNA molecule in the sample to provide tagged strands. In some embodiments, the method further comprises sequencing at least some of the tagged strands to produce a set of sequence reads. In some embodiments, the method further comprises sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a

first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads. In some embodiments, the method further comprises determining quantitative measures of (i) the paired reads and (ii) the unpaired reads that map to each of one or more genetic loci to determine a quantitative measure of total double-stranded DNA molecules in the sample that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads mapping to each locus.

[0062] In another aspect, a method for reducing distortion in a sequencing assay, comprises (a) tagging control parent polynucleotides with a first tag set to produce tagged control parent polynucleotides; (b) tagging test parent polynucleotides with a second tag set to produce tagged test parent polynucleotides; (c) mixing tagged control parent polynucleotides with tagged test parent polynucleotides to form a pool; (d) determining quantities of tagged control parent polynucleotides and tagged test parent polynucleotides; and (e) using the quantities of tagged control parent polynucleotides to reduce distortion in the quantities of tagged test parent polynucleotides.

[0063] In some embodiments, the first tag set comprises a plurality of tags, wherein each tag in the first tag set comprises a same control tag and an identifying tag, and wherein the first tag set comprises a plurality of different identifying tags. In some embodiments, the second tag set comprises a plurality of tags, wherein each tag in the second tag set comprises a same test tag and an identifying tag, wherein the test tag is distinguishable from the control tag, and wherein the second tag set comprises a plurality of different identifying tags. In some embodiments, (d) comprises amplifying tagged parent polynucleotides in the pool to form a pool of amplified, tagged polynucleotides, and sequencing amplified, tagged polynucleotides in the amplified pool to produce a plurality of sequence reads. In some embodiments, the method further comprises grouping sequence reads into families, each family comprising sequence reads generated from a same parent polynucleotide, which grouping is optionally based on information from an identifying tag and from start/end sequences of the parent polynucleotides, and, optionally, determining a consensus sequence for each of a plurality of parent polynucleotides from the plurality of sequence reads in a group.

[0064] In some embodiments, (d) comprises determining copy number variation in the test parent polynucleotides at greater than or equal to one locus based on relative quantity of test parent polynucleotides and control parent polynucleotides mapping to the locus.

[0065] Another aspect provides a method comprising (a) ligating adaptors to double-stranded DNA polynucleotides, wherein ligating is performed in a single reaction vessel, and wherein the adaptors comprise molecular barcodes, to produce a tagged library comprising an insert from the double-stranded DNA polynucleotides, and having between 4 and 1 million different tags; (b) generating a plurality of sequence reads for each of the double-stranded DNA polynucleotides in the tagged library; (c) grouping sequence reads into families, each family comprising sequence reads generated from a single DNA polynucleotide among the double-stranded DNA polynucleotides, based on information in a tag and information at an end of the insert; and (d) calling bases at each position in the double-stranded DNA molecule based on bases at the position in members of a family. In some embodiments, (b) comprises amplifying each of the double-stranded DNA polynucleotide molecules in the tagged library to generate amplification products, and sequencing the amplification products. In some embodiments, the method further comprises sequencing the double-stranded DNA polynucleotide molecules a plurality of times. In some embodiments, (b) comprises sequencing the entire insert. In some embodiments, (c) further comprises collapsing sequence reads in each family to generate a consensus sequence. In some embodiments, (d) comprises calling a plurality of sequential bases from at least a subset of the sequence reads to identify single nucleotide variations (SNV) in the double-stranded DNA molecule.

[0066] Another aspect provides a method of detecting disease cell heterogeneity from a sample comprising polynucleotides from somatic cells and disease cells. The method comprises quantifying polynucleotides in the sample bearing a nucleotide sequence variant at each of a plurality of genetic loci; determining copy number variation (CNV) at each of the plurality of genetic loci, wherein the CNV indicates a genetic dose of a locus in the disease cell polynucleotides; determining with a programmed computer processor a relative measure of quantity of polynucleotides bearing a sequence variant at a locus per the genetic dose at the locus for each of a plurality of the loci; and comparing the relative measures at each of the plurality of loci, wherein different relative measures is indicative of tumor heterogeneity.

[0067] In another aspect, a method comprises subjecting a subject to one or more pulsed therapy cycles, each pulsed therapy cycle comprising (a) a first period during which a drug is administered at a first amount; and (b) a second period during which the drug is administered at a second, reduced amount, wherein (i) the first period is characterized by a tumor burden detected above a first clinical level; and (ii) the second period is characterized by a tumor burden detected below a second clinical level.

[0068] Additional aspects and advantages of the present disclosure will become readily apparent to those skilled in this art from the following detailed description, wherein only illustrative embodiments of the present disclosure are shown and described. As will be realized, the present disclosure is capable of other and different embodiments, and its several details are capable of modifications in various obvious respects, all without departing from the disclosure. Accordingly, the drawings and description are to be regarded as illustrative in nature, and not as restrictive.

INCORPORATION BY REFERENCE

[0069] All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference.

BRIEF DESCRIPTION OF THE DRAWINGS

[0001] The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings (also "figure" and "FIG." herein), of which:

[0070] **FIG. 1** is a flowchart representation of a method of the present disclosure for determining copy number variation (CNV);

[0071] **FIG. 2** depicts mapping of pairs and singlets to Locus A and Locus B in a genome;

[0072] **FIG. 3** shows a reference sequence encoding a genetic Locus A;

[0073] **FIGs. 4A-C** shows amplification, sequencing, redundancy reduction and pairing of complementary molecules;

[0074] **FIG. 5** shows increased confidence in detecting sequence variants by pairing reads from Watson and Crick strands;

[0075] **FIG. 6** shows a computer system that is programmed or otherwise configured to implement various methods of the present disclosure;

[0076] **FIG. 7** is schematic representation of a system for analyzing a sample comprising nucleic acids from a user, including a sequencer; bioinformatic software and internet connection for report analysis by, for example, a hand held device or a desk top computer;

[0077] **FIG. 8** is a flowchart representation of a method of this invention for determining CNV using pooled test and control pools; and

[0078] **FIGs. 9A-9C** schematically illustrate a method for tagging a polynucleotide molecule with a library adaptor and subsequently a sequencing adaptor.

DETAILED DESCRIPTION

[0079] While various embodiments of the invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions may occur to those skilled in the art without departing from the invention. It should be understood that various alternatives to the embodiments of the invention described herein may be employed.

[0080] The term "genetic variant," as used herein, generally refers to an alteration, variant or polymorphism in a nucleic acid sample or genome of a subject. Such alteration, variant or polymorphism can be with respect to a reference genome, which may be a reference genome of the subject or other individual. Single nucleotide polymorphisms (SNPs) are a form of polymorphisms. In some examples, one or more polymorphisms comprise one or more single nucleotide variations (SNVs), insertions, deletions, repeats, small insertions, small deletions, small repeats, structural variant junctions, variable length tandem repeats, and/or flanking sequences,. Copy number variants (CNVs), transversions and other rearrangements are also forms of genetic variation. A genomic alternation may be a base change, insertion, deletion, repeat, copy number variation, or transversion.

[0081] The term "polynucleotide," as used herein, generally refers to a molecule comprising one or more nucleic acid subunits. A polynucleotide can include one or more subunits selected from adenosine (A), cytosine (C), guanine (G), thymine (T) and uracil (U), or variants thereof. A nucleotide can include A, C, G, T or U, or variants thereof. A nucleotide can include any subunit that can be incorporated into a growing nucleic acid strand. Such subunit can be an A, C, G, T, or U, or any other subunit that is specific to one or more complementary A, C, G, T or U, or complementary to a purine (i.e., A or G, or variant thereof) or a pyrimidine (i.e., C, T or U, or variant thereof). A subunit can enable individual nucleic acid bases or groups of bases (*e.g.*, AA, TA, AT, GC, CG, CT, TC, GT, TG, AC, CA, or uracil-counterparts thereof) to be resolved. In some examples, a polynucleotide is deoxyribonucleic acid (DNA) or ribonucleic acid (RNA), or derivatives thereof. A polynucleotide can be single-stranded or double stranded.

[0082] The term "subject," as used herein, generally refers to an animal, such as a mammalian species (*e.g.*, human) or avian (*e.g.*, bird) species, or other organism, such as a plant. More specifically, the subject can be a vertebrate, a mammal, a mouse, a primate, a simian or a human. Animals include, but are not limited to, farm animals, sport animals, and pets. A subject can be a healthy individual, an individual that has or is suspected of having a disease or a pre-disposition to the disease, or an individual that is in need of therapy or suspected of needing therapy. A subject can be a patient.

[0083] The term "genome" generally refers to an entirety of an organism's hereditary information. A genome can be encoded either in DNA or in RNA. A genome can comprise coding regions that code for proteins as well as non-coding regions. A genome can include the sequence of all chromosomes together in an organism. For example, the human genome has a total of 46 chromosomes. The sequence of all of these together constitutes a human genome.

[0084] The terms "adaptor(s)", "adapter(s)" and "tag(s)" are used synonymously throughout this specification. An adaptor or tag can be coupled to a polynucleotide sequence to be "tagged" by any approach including ligation, hybridization, or other approaches.

[0085] The term "library adaptor" or "library adapter" as used herein, generally refers to a molecule (e.g., polynucleotide) whose identity (e.g., sequence) can be used to differentiate polynucleotides in a biological sample (also "sample" herein).

[0086] The term "sequencing adaptor," as used herein, generally refers to a molecule (e.g., polynucleotide) that is adapted to permit a sequencing instrument to sequence a target polynucleotide, such as by interacting with the target polynucleotide to enable sequencing. The sequencing adaptor permits the target polynucleotide to be sequenced by the sequencing instrument. In an example, the sequencing adaptor comprises a nucleotide sequence that hybridizes or binds to a capture polynucleotide attached to a solid support of a sequencing system, such as a flow cell. In another example, the sequencing adaptor comprises a nucleotide sequence that hybridizes or binds to a polynucleotide to generate a hairpin loop, which permits the target polynucleotide to be sequenced by a sequencing system. The sequencing adaptor can include a sequencer motif, which can be a nucleotide sequence that is complementary to a flow cell sequence of other molecule (e.g., polynucleotide) and usable by the sequencing system to sequence the target polynucleotide. The sequencer motif can also include a primer sequence for use in sequencing, such as sequencing by synthesis. The

sequencer motif can include the sequence(s) needed to couple a library adaptor to a sequencing system and sequence the target polynucleotide.

[0087] As used herein the terms "at least", "at most" or "about", when preceding a series, refers to each member of the series, unless otherwise identified.

[0088] The term *"about"* and its grammatical equivalents in relation to a reference numerical value can include a range of values up to plus or minus 10% from that value. For example, the amount *"about 10"* can include amounts from 9 to 11. In other embodiments, the term *"about"* in relation to a reference numerical value can include a range of values plus or minus 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, or 1% from that value.

[0089] The term "at least" and its grammatical equivalents in relation to a reference numerical value can include the reference numerical value and greater than that value. For example, the amount "at least 10" can include the value 10 and any numerical value above 10, such as 11, 100, and 1,000.

[0090] The term "at most" and its grammatical equivalents in relation to a reference numerical value can include the reference numerical value and less than that value. For example, the amount "at most 10" can include the value 10 and any numerical value under 10, such as 9, 8, 5, 1, 0.5, and 0.1.

[0091] **1. Methods for processing and/or analyzing a nucleic acid sample**

[0092] An aspect of the present disclosure provides methods for determining a genomic alternation in a nucleic acid sample of a subject. **FIG. 1** shows a method of determining copy number variation (CNV). The method can be implemented to determine other genomic alternations, such as SNVs.

[0093] **A. Polynucleotide Isolation**

[0094] Methods disclosed herein can comprise isolating one or more polynucleotides. A polynucleotide can comprise any type of nucleic acid, for example, a sequence of genomic nucleic acid, or an artificial sequence (*e.g.*, a sequence not found in genomic nucleic acid). For example, an artificial sequence can contain non-natural nucleotides. Also, a polynucleotide can comprise both genomic nucleic acid and an artificial sequence, in any portion. For example, a polynucleotide can comprise 1 to 99% of genomic nucleic acid and 99% to 1% of artificial sequence, where the total adds up to 100%. Thus, fractions of percentages are also contemplated. For example, a ratio of 99.1% to 0.9% is contemplated.

[0095] A polynucleotide can comprise any type of nucleic acids, such as DNA and/or RNA. For example, if a polynucleotide is DNA, it can be genomic DNA, complementary

DNA (cDNA), or any other deoxyribonucleic acid. A polynucleotide can also be cell-free DNA (cfDNA). For example, the polynucleotide can be circulating DNA. The circulating DNA can comprise circulating tumor DNA (ctDNA). A polynucleotide can be double-stranded or single-stranded. Alternatively, a polynucleotide can comprise a combination of a double-stranded portion and a single-stranded portion.

[0096] Polynucleotides do not have to be cell-free. In some cases, the polynucleotides can be isolated from a sample. For example, in step (102) (**FIG. 1**), double-stranded polynucleotides are isolated from a sample. A sample can be any biological sample isolated from a subject. For example, a sample can comprise, without limitation, bodily fluid, whole blood, platelets, serum, plasma, stool, red blood cells, white blood cells or leucocytes, endothelial cells, tissue biopsies, synovial fluid, lymphatic fluid, ascites fluid, interstitial or extracellular fluid, the fluid in spaces between cells, including gingival crevicular fluid, bone marrow, cerebrospinal fluid, saliva, mucous, sputum, semen, sweat, urine, or any other bodily fluids. A bodily fluid can include saliva, blood, or serum. For example, a polynucleotide can be cell-free DNA isolated from a bodily fluid, *e.g.*, blood or serum. A sample can also be a tumor sample, which can be obtained from a subject by various approaches, including, but not limited to, venipuncture, excretion, ejaculation, massage, biopsy, needle aspirate, lavage, scraping, surgical incision, or intervention or other approaches.

[0097] A sample can comprise various amount of nucleic acid that contains genome equivalents. For example, a sample of about 30 ng DNA can contain about 10,000 (10^4) haploid human genome equivalents and, in the case of cfDNA, about 200 billion ($2x10^{11}$) individual polynucleotide molecules. Similarly, a sample of about 100 ng of DNA can contain about 30,000 haploid human genome equivalents and, in the case of cfDNA, about 600 billion individual molecules.

[0098] A sample can comprise nucleic acids from different sources. For example, a sample can comprise germline DNA or somatic DNA. A sample can comprise nucleic acids carrying mutations. For example, a sample can comprise DNA carrying germline mutations and/orsomatic mutations, . A sample can also comprise DNA carrying cancer-associated mutations (*e.g.*, cancer-associated somatic mutations).

[0099] **B. <u>Tagging</u>**

[00100] Polynucleotides disclosed herein can be tagged. For example, in step (104) (**FIG. 1**) the double-stranded polynucleotides are tagged with duplex tags, tags that differently label the complementary strands (*i.e.*, the "Watson" and "Crick" strands) of a double-stranded

molecule. In one embodiment the duplex tags are polynucleotides having complementary and non-complementary portions.

[00101] Tags can be any types of molecules attached to a polynucleotide, including, but not limited to, nucleic acids, chemical compounds, florescent probes, or radioactive probes. Tags can also be oligonucleotides (*e.g.*, DNA or RNA). Tags can comprise known sequences, unknown sequences, or both. A tag can comprise random sequences, pre-determined sequences, or both. A tag can be double-stranded or single-stranded. A double-stranded tag can be a duplex tag. A double-stranded tag can comprise two complementary strands. Alternatively, a double-stranded tag can comprise a hybridized portion and a non-hybridized portion. The double-stranded tag can be Y-shaped, *e.g.*, the hybridized portion is at one end of the tag and the non- hybridized portion is at the opposite end of the tag. One such example are the "Y adapters" used in Illumina sequencing. Other examples include hairpin shaped adapters or bubble shaped adapters. Bubble shaped adapters have non-complementary sequences flanked on both sides by complementary sequences.

[00102] Tagging disclosed herein can be performed using any method. A polynucleotide can be tagged with an adaptor by hybridization. For example, the adaptor can have a nucleotide sequence that is complementary to at least a portion of a sequence of the polynucleotide. As an alternative, a polynucleotide can be tagged with an adaptor by ligation.

[00103] For example, tagging can comprise using one or more enzymes. The enzyme can be a ligase. The ligase can be a DNA ligase. For example, the DNA ligase can be a T4 DNA ligase, *E. coli* DNA ligase, and/or mammalian ligase. The mammalian ligase can be DNA ligase I, DNA ligase III, or DNA ligase IV. The ligase can also be a thermostable ligase. Tags can be ligated to a blunt-end of a polynucleotide (blunt-end ligation). Alternatively, tags can be ligated to a sticky end of a polynucleotide (sticky-end ligation). Efficiency of ligation can be increased by optimizing various conditions. Efficiency of ligation can be increased by optimizing the reaction time of ligation. For example, the reaction time of ligation can be less than 12 hours, *e.g.*, less than 1, less than 2, less than 3, less than 4, less than 5, less than 6, less than 7, less than 8, less than 9, less than 10, less than 11, less than 12, less than 13, less than 14, less than 15, less than 16, less than 17, less than 18, less than 19, or less than 20 hours. In a particular example, reaction time of ligation is less than 20 hours. Efficiency of ligation can be increased by optimizing the ligase concentration in the reaction. For example, the ligase concentration can be at least 10, at least 50, at least 100, at least 150,

at least 200, at least 250, at least 300, at least 400, at least 500, or at least 600 unit/microliter. Efficiency can also be optimized by adding or varying the concentration of an enzyme suitable for ligation, enzyme cofactors or other additives, and/or optimizing a temperature of a solution having the enzyme. Efficiency can also be optimized by varying the addition order of various components of the reaction. The end of tag sequence can comprise dinucleotide to increase ligation efficiency. When the tag comprises a non-complementary portion (e.g., Y-shaped adaptor), the sequence on the complementary portion of the tag adaptor can comprise one or more selected sequences that promote ligation efficiency. Preferably such sequences are located at the terminal end of the tag. Such sequences can comprise 1, 2, 3, 4, 5, or 6 terminal bases. Reaction solution with high viscosity (*e.g.*, a low Reynolds number) can also be used to increase ligation efficiency. For example, solution can have a Reynolds number less than 3000, less than 2000, less than 1000, less than 900, less than 800, less than 700, less than 600, less than 500, less than 400, less than 300, less than 200, less than 100, less than 50, less than 25, or less than 10. It is also contemplated that roughly unified distribution of fragments (*e.g.*, tight standard deviation) can be used to increase ligation efficiency. For example, the variation in fragment sizes can vary by less than 20%, less than 15%, less than 10%, less than 5%, or less than 1%. Tagging can also comprise primer extension, for example, by polymerase chain reaction (PCR). Tagging can also comprise any of ligation-based PCR, multiplex PCR, single strand ligation, or single strand circularization.

[00104] In some instances, the tags herein comprise molecular barcodes. Such molecular barcodes can be used to differentiate polynucleotides in a sample. Preferably molecular barcodes are different from one another. For example, molecular barcodes can have a difference between them that can be characterized by a predetermined edit distance or a Hamming distance. In some instances, the molecular barcodes herein have a minimum edit distance of 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10. To further improve efficiency of conversion (e.g., tagging) of untagged molecular to tagged molecules, one preferably utilizes short tags. For example, in some embodiments, a library adapter tag can be up to 65, 60, 55, 50, 45, 40, or 35 nucleotide bases in length. A collection of such short library barcodes preferably includes a number of different molecular barcodes, e.g., at least 2, 4, 6, 8, 10, 12, 14, 16, 18 or 20 different barcodes with a minimum edit distance of 1, 2, 3 or more.

[00105] Thus, a collection of molecules can include one or more tags. In some instances, some molecules in a collection can include an identifying tag ("identifier") such as a molecular barcode that is not shared by any other molecule in the collection. For example, in

some instances of a collection of molecules, at least 50%, at least 51%, at least 52%, at least 53%, at least 54%, at least 55%, at least 56%, at least 57%, at least 58%, at least 59%, at least 60%, at least 61%, at least 62%, at least 63%, at least 64%, at least 65%, at least 66%, at least 67%, at least 68%, at least 69%, at least 70%, at least 71%, at least 72%, at least 73%, at least 74%, at least 75%, at least 76%, at least 77%, at least 78%, at least 79%, at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% of the molecules in the collection can include an identifier or molecular barcode that is not shared by any other molecule in the collection. As used herein, a collection of molecules is considered to be "uniquely tagged" if each of at least 95% of the molecules in the collection bears an identifier that is not shared by any other molecule in the collection ("unique tag" or "unique identifier"). A collection of molecules is considered to be "non-uniquely tagged" if each of at least 1%, at least 5%, at least 10%, at least 15%, at least 20%, at least 25%, at least 30%, at least 35%, at least 40%, at least 45%, or at least or about 50% of the molecules in the collection bears an identifying tag or molecular barcode that is shared by at least one other molecule in the collection ("non-unique tag" or "non-unique identifier"). Accordingly, in a non-uniquely tagged population no more than 1% of the molecules are uniquely tagged. For example, in a non-uniquely tagged population, no more than 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, or 50% of the molecules can be uniquely tagged.

[00106] A number of different tags can be used based on the estimated number of molecules in a sample. In some tagging methods, the number of different tags can be at least the same as the estimated number of molecules in the sample. In other tagging methods, the number of different tags can be at least two, three, four, five, six, seven, eight, nine, ten, one hundred or one thousand times as many as the estimated number of molecules in the sample. In unique tagging, at least two times (or more) as many different tags can be used as the estimated number of molecules in the sample.

[00107] The molecules in the sample may be non-uniquely tagged. In such instances a fewer number of tags or molecular barcodes is used then the number of molecules in the sample to be tagged. For example, no more than 100, 50, 40, 30, 20 or 10 unique tags or molecular barcodes are used to tag a complex sample such as a cell free DNA sample with many more different fragments.

[00108] The polynucleotide to be tagged can be fragmented, such as either naturally or using other approaches, such as, for example, shearing. The polynucleotides can be fragmented by certain methods, including but not limited to, mechanical shearing, passing the sample through a syringe, sonication, heat treatment (*e.g.*, for 30 minutes at 90°C), and/or nuclease treatment (*e.g.*, using DNase, RNase, endonuclease, exonuclease, and/or restriction enzyme).

[00109] The polynucleotides fragments (prior to tagging) can comprise sequences of any length. For example, polynucleotide fragments (prior to tagging) can comprise at least 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, 145, 150, 155, 160, 165, 170, 175, 180, 185, 190, 195, 200, 205, 210, 215, 220, 225, 230, 235, 240, 245, 250, 255, 260, 265, 270, 275, 280, 285, 290, 295, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000 or more nucleotides in length. The polynucleotide fragment are preferably about the average length of cell-free DNA. For example, the polynucleotide fragments can comprise about 160 bases in length. The polynucleotide fragment can also be fragmented from a larger fragment into smaller fragments about 160 bases in length.

[00110] Polynucleotides tagged can comprise sequences associated with cancer. The cancer-associated sequences can comprise single nucleotide variation (SNV), copy number variation (CNV), insertions, deletions, and/or rearrangements.

[00111] The polynucleotides can comprise sequences associated with cancer, such as acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), adrenocortical carcinoma, Kaposi Sarcoma, anal cancer, basal cell carcinoma, bile duct cancer, bladder cancer, bone cancer, osteosarcoma, malignant fibrous histiocytoma, brain stem glioma, brain cancer, craniopharyngioma, ependymoblastoma, ependymoma, medulloblastoma, medulloeptithelioma, pineal parenchymal tumor, breast cancer, bronchial tumor, Burkitt lymphoma, Non-Hodgkin lymphoma, carcinoid tumor, cervical cancer, chordoma, chronic lymphocytic leukemia (CLL), chronic myelogenous leukemia (CML), colon cancer, colorectal cancer, cutaneous T-cell lymphoma, ductal carcinoma in situ, endometrial cancer, esophageal cancer, Ewing Sarcoma, eye cancer, intraocular melanoma, retinoblastoma, fibrous histiocytoma, gallbladder cancer, gastric cancer, glioma, hairy cell leukemia, head and neck cancer, heart cancer, hepatocellular (liver) cancer, Hodgkin lymphoma, hypopharyngeal cancer, kidney cancer, laryngeal cancer, lip cancer, oral cavity cancer, lung cancer, non-small cell carcinoma, small cell carcinoma, melanoma, mouth cancer,

myelodysplastic syndromes, multiple myeloma, medulloblastoma, nasal cavity cancer, paranasal sinus cancer, neuroblastoma, nasopharyngeal cancer, oral cancer, oropharyngeal cancer, osteosarcoma, ovarian cancer, pancreatic cancer, papillomatosis, paraganglioma, parathyroid cancer, penile cancer, pharyngeal cancer, pituitary tumor, plasma cell neoplasm, prostate cancer, rectal cancer, renal cell cancer, rhabdomyosarcoma, salivary gland cancer, Sezary syndrome, skin cancer, nonmelanoma, small intestine cancer, soft tissue sarcoma, squamous cell carcinoma, testicular cancer, throat cancer, thymoma, thyroid cancer, urethral cancer, uterine cancer, uterine sarcoma, vaginal cancer, vulvar cancer, Waldenstrom macroglobulinemia, and/or Wilms Tumor.

[00112] A haploid human genome equivalent has about 3 picograms of DNA. A sample of about 1 microgram of DNA contains about 300,000 haploid human genome equivalents. Improvements in sequencing can be achieved as long as at least some of the duplicate or cognate polynucleotides bear unique identifiers with respect to each other, that is, bear different tags. However, in certain embodiments, the number of tags used is selected so that there is at least a 95% chance that all duplicate molecules starting at any one position bear unique identifiers. For example, in a sample comprising about 10,000 haploid human genome equivalents of fragmented genomic DNA, *e.g.*, cfDNA, z is expected to be between 2 and 8. Such a population can be tagged with between about 10 and 100 different identifiers, for example, about 2 identifiers, about 4 identifiers, about 9 identifiers, about 16 identifiers, about 25 identifiers, about 36 different identifiers, about 49 different identifiers, about 64 different identifiers, about 81 different identifiers, or about 100 different identifiers.

[00113] Nucleic acid barcodes having identifiable sequences including molecular barcodes, can be used for tagging. For example, a plurality of DNA barcodes can comprise various numbers of sequences of nucleotides. A plurality of DNA barcodes having 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more identifiable sequences of nucleotides can be used. When attached to only one end of a polynucleotide, the plurality of DNA barcodes can produce 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more different identifiers. Alternatively, when attached to both ends of a polynucleotide, the plurality DNA barcodes can produce 4, 9, 16, 25, 36, 49, 64, 81, 100, 121, 144, 169, 196, 225, 256, 289, 324, 361, 400 or more different identifiers (which is the ^2 of when the DNA barcode is attached to only 1 end of a polynucleotide). In one example, a plurality of DNA barcodes having 6, 7, 8, 9 or 10 identifiable sequences of nucleotides can be used. When attached to

both ends of a polynucleotide, they produce 36, 49, 64, 81 or 100 possible different identifiers, respectively. In a particular example, the plurality of DNA barcodes can comprise 8 identifiable sequences of nucleotides. When attached to only one end of a polynucleotide, the plurality of DNA barcodes can produce 8 different identifiers. Alternatively, when attached to both ends of a polynucleotide, the plurality of DNA barcodes can produce 64 different identifiers. Samples tagged in such a way can be those with a range of about 10 ng to any of about 100 ng, about 1 μg, about 10 μg of fragmented polynucleotides, *e.g.*, genomic DNA, *e.g.*, cfDNA.

[00114] A polynucleotide can be uniquely identified in various ways. A polynucleotide can be uniquely identified by a unique DNA barcode. For example, any two polynucleotides in a sample are attached two different DNA barcodes. Alternatively, a polynucleotide can be uniquely identified by the combination of a DNA barcode and one or more endogenous sequences of the polynucleotide. For example, any two polynucleotides in a sample can be attached the same DNA barcode, but the two polynucleotides can still be identified by different endogenous sequences. The endogenous sequence can be on an end of a polynucleotide. For example, the endogenous sequence can be adjacent (*e.g.*, base in between) to the attached DNA barcode. In some instances the endogenous sequence can be at least 2, 4, 6, 8, 10, 20, 30, 40, 50, 60, 70, 80, 90, or 100 bases in length. Preferably, the endogenous sequence is a terminal sequence of the fragment/polynucleotides to be analyzed. The endogenous sequence may be the length of the sequence. For example, a plurality of DNA barcodes comprising 8 different DNA barcodes can be attached to both ends of each polynucleotide in a sample. Each polynucleotide in the sample can be identified by the combination of the DNA barcodes and about 10 base pair endogenous sequence on an end of the polynucleotide. Without being bound by theory, the endogenous sequence of a polynucleotide can also be the entire polynucleotide sequence.

[00115] Also disclosed herein are compositions of tagged polynucleotides. The tagged polynucleotide can be single-stranded. Alternatively, the tagged polynucleotide can be double-stranded (*e.g.*, duplex-tagged polynucleotides). Accordingly, this invention also provides compositions of duplex-tagged polynucleotides. The polynucleotides can comprise any types of nucleic acids (DNA and/or RNA). The polynucleotides comprise any types of DNA disclosed herein. For example, the polynucleotides can comprise DNA, *e.g.*, fragmented DNA or cfDNA. A set of polynucleotides in the composition that map to a mappable base position in a genome can be non-uniquely tagged, that is, the number of

different identifiers can be at least 2 and fewer than the number of polynucleotides that map to the mappable base position. The number of different identifiers can also be at least 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25and fewer than the number of polynucleotides that map to the mappable base position.

[00116] In some instances, as a composition goes from about 1 ng to about 10 μg or higher, a larger set of different molecular barcodes can be used. For example, between 5 and 100 different library adaptors can be used to tag polynucleotides in a cfDNA sample.

[00117] The systems and methods disclosed herein may be used in applications that involve the assignment of molecular barcodes. The molecular barcodes can be assigned to any types of polynucleotides disclosed in this invention. For example, the molecular barcodes can be assigned to cell-free polynucleotides (*e.g.*, cfDNAs). Often, an identifier disclosed herein can be a barcode oligonucleotide that is used to tag the polynucleotide. The barcode identifier may be a nucleic acid oligonucleotide (*e.g.*, a DNA oligonucleotide). The barcode identifier can be single-stranded. Alternatively, the barcode identifier can be double-stranded. The barcode identifier can be attached to polynucleotides using any method disclosed herein. For example, the barcode identifier can be attached to the polynucleotide by ligation using an enzyme. The barcode identifier can also be incorporated into the polynucleotide 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, for example, U.S. patent applications 2001/0053519, 2003/0152490, 2011/0160078 and U.S. Patent No. 6,582,908, each of which is entirely incorporated herein by reference.

[00118] Identifiers or molecular barcodes used herein may be completely endogenous whereby circular ligation of individual fragments may be performed followed by random shearing or targeted amplification. In this case, the combination of a new start and stop point of the molecule and the original intramolecular ligation point can form a specific identifier.

[00119] Identifiers or molecular barcodes used herein can comprise any types of oligonucleotides. In some cases, identifiers may be predetermined, random, or semi-random sequence oligonucleotides. Identifiers can be barcodes. For example, a plurality of barcodes may be used such that barcodes are not necessarily unique to one another in the plurality. Alternatively, a plurality of barcodes may be used such that each barcode is unique to any other barcode in the plurality. The barcodes can comprise specific sequences (*e.g.*,

predetermined sequences) that can be individually tracked. Further, barcodes may be attached (*e.g.*, by ligation) to individual molecules such that the combination of the barcode and the sequence it may be ligated to creates a specific sequence that may be individually tracked. As described herein, detection of barcodes in combination with sequence data of beginning (start) and/or end (stop) portions of sequence reads can 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. In this way the polynucleotides in the sample can be uniquely or substantially uniquely tagged. A duplex tag can include a degenerate or semi-degenerate nucleotide sequence, *e.g.*, a random degenerate sequence. The nucleotide sequence can comprise any number of nucleotides. For example, the nucleotide sequence can comprise 1 (if using a non-natural nucleotide), 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 or more nucleotides. In a particular example, the sequence can comprise 7 nucleotides. In another example, the sequence can comprise 8 nucleotides. The sequence can also comprise 9 nucleotides. The sequence can comprise 10 nucleotides.

[00120] A barcode can comprise contiguous or non-contiguous sequences. A barcode that comprises at least 1, 2, 3, 4, 5 or more nucleotides is a contiguous sequence or non-contiguous sequence. if the 4 nucleotides are uninterrupted by any other nucleotide. For example, if a barcode comprises the sequence TTGC, a barcode is contiguous if the barcode is TTGC. On the other hand, a barcode is non-contiguous if the barcode is TTXGC, where X is a nucleic acid base.

[00121] An identifier or molecular barcode can have an n-mer sequence which may be 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 or more nucleotides in length. A tag herein can comprise any range of nucleotides in length. For example, the sequence can be between 2 to 100, 10 to 90, 20 to 80, 30 to 70, 40 to 60, or about 50 nucleotides in length.

[00122] The tag can comprise a double-stranded fixed reference sequence downstream of the identifier or molecular barcode. Alternatively, the tag can comprise a double-stranded fixed reference sequence upstream or downstream of the identifier or molecular barcode.

Each strand of a double-stranded fixed reference sequence can be, for example, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 nucleotides in length.

[00123] **C. Adaptors**

[00124] A library of polynucleotide molecules can be synthesized for use in sequencing. For example, a library of polynucleotides comprising a plurality of polynucleotide molecules that are each less than or equal to 100, 90, 80, 70, 60, 50, 45, 40, or 35 nucleic acid (or nucleotide) bases in length can be made. A plurality of polynucleotide molecules can be each less than or equal to 35 nucleic acid bases in length. A plurality of polynucleotide molecules can be each less than or equal to 30 nucleic acid bases in length. A plurality of polynucleotide molecules can also be less than or equal to 250, 200, 150, 100, or 50 nucleic acid bases. Additionally, the plurality of polynucleotide molecules can also be less than or equal to 100, 99, 98, 97, 96, 95, 94, 93, 92, 91, 90, 89, 88, 87, 86, 85, 84, 83, 82, 81, 80, 79, 78, 77, 76, 75, 74, 73, 72, 71, 70, 69, 68, 67, 66, 65, 64, 63, 62, 61, 60, 59, 58, 57, 56, 55, 54, 53, 52, 51, 50, 49, 48, 47, 46, 45, 44, 43, 42, 41, 40, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, or 10 nucleic acid bases.

[00125] A library of polynucleotides comprising a plurality of polynucleotide molecules can also have distinct (with respect to each other) molecular barcode sequences (or molecular barcodes) with respect to at least 4 nucleic acid bases. A molecular barcode (also "barcode" or "identifier" herein) sequence is a nucleotide sequence that distinguishes one polynucleotide from another. In other embodiments, the polynucleotide molecules can also have different barcode sequences with respect to 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 or more nucleic acid bases.

[00126] A library of polynucleotides comprising a plurality of polynucleotide molecules can also have a plurality of different barcode sequences. For example, a plurality of polynucleotide molecules can have at least 4 different molecular barcode sequences. In some cases, the plurality of polynucleotide molecules has from 2-100, 4-50, 4-30, 4-20, or 4-10 different molecular barcode sequences. The plurality of polynucleotides molecules can also have other ranges of different barcode sequences such as, 1-4, 2-5, 3-6, 4-7, 5-8, 6-9, 7-10, 8-11, 9-12, 10-13, 11-14, 12-15, 13-16, 14-17, 15-18, 16-19, 17-20, 18-21, 19-22, 20-23, 21-24, or 22-25 different barcode sequences. In other cases, a plurality of polynucleotide molecules can have at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24,

25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, or 100 more different barcode sequences. In a particular example, the plurality library adapters comprise at least 8 different sequences.

[00127] The location of the different barcode sequences can vary within the plurality of polynucleotides. For example, the different barcode sequences can be within 20, 15, 10, 9, 8, 7, 6, 5, 4, 3, or 2 nucleic acid bases from a terminal end of a respective one of the plurality of polynucleotide molecules. In an example, a plurality of polynucleotide molecules has distinct barcode sequences that are within 10 nucleic acid bases from the terminal end. In another example, a plurality of polynucleotide molecules has distinct barcode sequences that are within 5 or 1 nucleic acid bases from the terminal end. In other instances, the distinct barcode sequences can be at the terminal end of a respective one of the plurality of polynucleotide molecules. Other variations include that the distinct molecular barcode sequences can be within 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, or 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, or more nucleic acid bases from a terminal end of a respective one of the plurality of polynucleotide molecules.

[00128] The terminal end of the plurality of polynucleotide molecules can be adapted for ligation to a target nucleic acid molecule. For example, the terminal end can be a blunt end. In some other cases, the terminal end is adapted for hybridization to a complementary sequence of a target nucleic acid molecule.

[00129] A library of polynucleotides comprising a plurality of polynucleotide molecules can also have an edit distance of at least 1. In some cases, the edit distance is with respect to individual bases of the plurality of polynucleotide molecules. In other cases, the plurality of polynucleotide molecules can have an edit distance of at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,

12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 or more. The edit distance can be a Hamming distance.

[00130] In some cases, the plurality of polynucleotides does not contain sequencing adaptors. A sequence adaptor can be a polynucleotide that comprises a sequence that hybridizes to one or more sequencing adaptors or primers. A sequencing adaptor can further comprise a sequence hybridizing to a solid support, *e.g.*, a flow cell sequence. The term "flow cell sequence" and its grammatical equivalents as used herein, refers to a sequence that permits hybridization to a substrate, for example, by way of a primer attached to the substrate. The substrate can be bead or a planar surface. In some embodiments, a flow cell sequence can allow a polynucleotide to attach to a flow cell or surface (e.g., surface of a bead, for example, an Illumina flow cell.

[00131] When a plurality of polynucleotide molecules does not contain sequencing adaptors or primers, each polynucleotide molecule of the plurality does not contain a nucleic acid sequence or other moiety that is adapted to permit sequencing of a target nucleic acid molecule with a given sequencing approach, such as Illumina, SOLiD, Pacific Biosciences, GeneReader, Oxford Nanopore, Complete Genomics, Gnu-Bio, Ion Torrent, Oxford Nanopore or Genia. In some examples, when a plurality of polynucleotide molecules does not contain sequencing adaptors or primers, the plurality of polynucleotide molecules does not contain flow cell sequences. For example, the plurality of polynucleotide molecules cannot bind to flow cells, such as used in Illumina flow cell sequencers. However, these flow cell sequences, if desired, can be added to the plurality of polynucleotide molecules by methods such as PCR amplification or ligation. At this point, Illumina flow cell sequencers can be used. Alternatively, when the plurality of polynucleotide molecules does not contain sequencing adaptors or primers, the plurality of polynucleotide molecules does not contain hairpin shaped adaptors or adaptors for generating hairpin loops in a target nucleic acid molecule, such as Pacific Bioscience SMRTbell™ adaptors. However, these hairpin shaped adaptors, if desired, can be added to the plurality of polynucleotide molecules by methods such as PCR amplification or ligation. The plurality of polynucleotide molecules can be circular or linear.

[00132] A plurality of polynucleotide molecules can be double stranded. In some cases, the plurality of polynucleotide molecules can be single stranded, or can comprise hybridized

and non-hybridized regions. A plurality of polynucleotide molecules can be non-naturally occurring polynucleotide molecules.

[00133] Adaptors can be polynucleotide molecules. The polynucleotide molecules can be Y-shaped, bubble-shaped or hairpin-shaped. A hairpin adaptor may contain a restriction site(s) or a Uracil containing base. Adaptors can comprise a complementary portion and a non-complementary portion. The non-complementary portion can have an edit distance (*e.g.*, Hamming distance). For example, the edit distance can be at least 1, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 21, at least 22, at least 23, at least 24, at least 25, at least 26, at least 27, at least 28, at least 29, or at least 30. The complementary portion of the adaptor can comprise sequences that are selected to enable and/or promote ligation to a polynucleotide, *e.g.*, a sequence to enable and/or promote ligation to a polynucleotide at a high yield.

[00134] A plurality of polynucleotide molecules as disclosed herein can be purified. In some cases, a plurality of polynucleotide molecules as disclosed herein can be isolated polynucleotide molecules. In other cases, a plurality of polynucleotide molecules as disclosed herein can be purified and isolated polynucleotide molecules.

[00135] In certain aspects, each of the plurality of polynucleotide molecules is Y-shaped or hairpin-shaped. Each of the plurality of polynucleotide molecules can comprise a different barcode. The different barcode can be a randomer in the complementary portion (*e.g.*, double stranded portion) of the Y-shaped or hairpin-shaped adaptor. Alternatively, the different barcode can be in one strand of the non-complementary portion (*e.g.*, one of the Y-shaped arms). As discussed above, the different barcode can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or more (or any length as described throughout) nucleic acid bases, *e.g.*, 7 bases. The barcode can be contiguous or non-contiguous sequences, as described above. The plurality of polynucleotide molecules is from 10 nucleic acid bases to 35 nucleic acid bases (or any length as described above) in length. Further, the plurality of polynucleotide molecules can comprise an edit distance (as described above), that is a Hamming distance. A plurality of polynucleotide molecules can have distinct barcode sequences that are within 10 nucleic acid bases from the terminal end.

[00136] In another aspect, a plurality of polynucleotide molecules can be sequencing adaptors. A sequencing adaptor can comprise a sequence hybridizing to one or more sequencing primers. A sequencing adaptor can further comprise a sequence hybridizing to a

solid support, *e.g.*, a flow cell sequence. For example, a sequencing adaptor can be a flow cell adaptor. The sequencing adaptors can be attached to one or both ends of a polynucleotide fragment. In another example, a sequencing adaptor can be hairpin shaped. For example, the hairpin shaped adaptor can comprise a complementary double-stranded portion and a loop portion, where the double-stranded portion can be attached (*e.g.*, ligated) to a double-stranded polynucleotide. Hairpin shaped sequencing adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times. A sequencing adaptor can be up to 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, or more bases from end to end. For example, a sequencing adaptor can be up to 70 bases from end to end. The sequencing adaptor can comprise 20-30, 20-40, 30-50, 30-60, 40-60, 40-70, 50-60, 50-70, bases from end to end. In a particular example, the sequencing adaptor can comprise 20-30 bases from end to end. In another example, the sequencing adaptor can comprise 50-60 bases from end to end. A sequencing adaptor can comprise one or more barcodes. For example, a sequencing adaptor can comprise a sample barcode. The sample barcode can comprise a pre-determined sequence. The sample barcodes can be used to identify the source of the polynucleotides. The sample barcode can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or more (or any length as described throughout) nucleic acid bases, *e.g.*, at least 8 bases. The barcode can be contiguous or non-contiguous sequences, as described above.

[00137] The plurality of polynucleotide molecules as described herein can be used as adaptors. Adaptors can comprise one or more identifiers. An adaptor can comprise an identifier with a random sequence. Alternatively, an adaptor can comprise an identifier with pre-determined sequences. Some adaptors can comprise an identifier with a random sequence and another identifier with a pre-determined sequence. The adaptors comprising identifiers can be double-stranded or single-stranded adaptors. The adaptors comprising identifiers can be Y-shaped adaptors. A Y-shaped adaptor can comprise one or more identifiers with a random sequence. The one or more identifiers can be on the hybrid portion and/or non-hybridized portion of the Y-shaped adaptor. A Y-shaped adaptor can comprise one or more identifiers with a pre-determined sequence. The one or more identifiers with pre-determined sequence can be on the hybridized portion and/or non- hybridized portion of

the Y-shaped adaptor. A Y-shaped adaptor can comprise one or more identifiers with a random sequence and one or more identifiers with a pre-determined sequence. For example, the one or more identifiers with a random sequence can be on the hybridized portion of the Y-shaped adaptor and/or the non- hybridized portion of the Y-shaped adaptor. The one or more identifiers with a pre-determined sequence can be on the hybridized portion of the Y-shaped adaptor and/or the non- hybridized portion of the Y-shaped adaptor. In a particular example, a Y-shaped adaptor can comprise an identifier with a random sequence on its hybridized portion and an identifier with a pre-determined sequence on its non- hybridized portion. The identifiers can be in any length disclosed herein. For example, a Y-shaped adaptor can comprise an identifier with a random sequence of 7 nucleotides on its hybridized portion and an identifier with a pre-determined sequence of 8 nucleotides on its non-hybridized portion.

[00138] An adaptor can include a double-stranded portion with a molecular barcode and at least one or two single-stranded portion. For example, the adaptor can be Y-shaped and include a double-stranded portion and two single-stranded portions. The single-stranded portions can include sequences that are not complementary to one another.

[00139] The adaptor can include a terminal end that has a sequence that is selected to permit the adaptor to be efficiently (e.g., at an efficiency of at least about 20%, 30%, 40%, 50%) ligated or otherwise coupled to a polynucleotide. In some examples, terminal nucleotides in a double-stranded portion of an adaptor are selected from a combination of purines and pyrimidines to provide for efficient ligation.

[00140] In some examples, a set of library adaptors comprises a plurality of polynucleotide molecules (library adaptors) with molecular barcodes. The library adaptors are less than or equal to 80, 70, 60, 50, 45, or 40 nucleotide bases in length. The molecular barcodes can be at least 4 nucleotide bases in length, but may be from 4 to 20 nucleotide bases in length. The molecular barcodes can be different from one another and have an edit distance of at least 1, 2, 3, 4, or 5 between one another. The molecular barcodes are located at least 1, 2, 3, 4, 5, 10, or 20 nucleotide bases away from a terminal end of their respective library adaptors. In some cases, the at least one terminal base is identical in all of the library adaptors.

[00141] The library adaptors can be identical but for the molecular barcodes. For example, the library adaptors can have identical sequences but differ only with respect to nucleotide sequences of the molecular barcodes.

[00142] Each of the library adaptors can have a double stranded portion and at least one single-stranded portion. By "single stranded portion" is meant an area of non-complementarity or an overhang. In some cases, each of the library adaptors has a double-stranded portion and two single-stranded portions. The double-stranded portion can have a molecular barcode. In some cases, the molecular barcode is a randomer. Each of the library adaptors can further include a strand-identification barcode on a single-stranded portion. The strand-identification barcode can include at least 4 nucleotide bases, in some cases from 4 to 20 nucleotide bases.

[00143] In some examples, each of the library adaptors has a double-stranded portion with a molecular barcode and two single-stranded portions. The single-stranded portions may not hybridize to one another. The single-stranded portions may not be completely complementary to one another.

[00144] The library adaptors can have a sequence of terminal nucleotides in a double-stranded portion that are the same. The sequence of terminal nucleotides can be at least 2, 3, 4, 5 or 6 nucleotide bases in length. For example, one strand of a double-stranded portion of the library adaptor can have the sequence ACTT, TCGC, or TACC at the terminal end, while the other strand can have a complementary sequence. In some cases, such a sequence is selected to optimize the efficiency at which the library adaptors ligate to target polynucleotides. Such sequences can be selected to optimize a binding interaction between the ends of the library adaptors and the target polynucleotides.

[00145] In some cases, none of the library adaptors contains a sample identification motif (or sample molecular barcode). Such sample identification motif can be provided via sequencing adaptors. A sample identification motif can include a sequencer of at least 4, 5, 6, 7, 8, 9, 10, 20, 30, or 40 nucleotide bases that permits the identification of polynucleotide molecules from a given sample from polynucleotide molecules from other samples. For example, this can permit polynucleotide molecules from two subjects to be sequenced in the same pool and sequence reads for the subjects subsequently identified.

[00146] A sequencer motif includes nucleotide sequence(s) needed to couple a library adaptor to a sequencing system and sequence a target polynucleotide coupled to the library adaptor. The sequencer motif can include a sequence that is complementary to a flow cell sequence and a sequence (sequencing initiation sequence) that is selectively hybridizable to a primer (or priming sequence) for use in sequencing. For example, such sequencing initiation sequence can be complementary to a primer that is employed for use in sequence by synthesis

(e.g., Illumina). Such primer can be included in a sequencing adaptor. A sequencing initiation sequence can be a primer hybridization site.

[00147] In some cases, none of the library adaptors contains a complete sequencer motif. The library adaptors can contain partial or no sequencer motifs. In some cases, the library adaptors include a sequencing initiation sequence. The library adaptors can include a sequencing initiation sequence but no flow cell sequence. The sequence initiation sequence can be complementary to a primer for sequencing. The primer can be a sequence specific primer or a universal primer. Such sequencing initiation sequences may be situated on single-stranded portions of the library adaptors. As an alternative, such sequencing initiation sequences may be priming sites (e.g., kinks or nicks) to permit a polymerase to couple to the library adaptors during sequencing.

[00148] In some cases, partial or complete sequencer motifs are provided by sequencing adaptors. A sequencing adaptor can include a sample molecular barcode and a sequencer motif. The sequencing adaptors can be provided in a set that is separate from the library adaptors. The sequencing adaptors in a given set can be identical – i.e., they contain the same sample barcode and sequencer motif.

[00149] Sequencing adaptors can include sample identification motifs and sequencer motifs. Sequencer motifs can include primers that are complementary to a sequencing initiation sequence. In some cases, sequencer motifs also include flow cell sequences or other sequences that permit a polynucleotide to a configured or arranged in a manner that permits the polynucleotide to be sequenced by a sequencer.

[00150] Library adaptors and sequencing adaptors can each be partial adaptors, that is, containing part but not all of the sequences necessary to enable sequencing by a sequencing platform. Together they provide complete adaptors. For example, library adaptors can include partial or no sequencer motifs, but such sequencer motifs are provided by sequencing adaptors.

[00151] **FIGs. 9A-9C** schematically illustrate a method for tagging a target polynucleotide molecule with library adaptors. **FIG. 9A** shows a library adaptor as a partial adaptor containing a primer hybridization site on one of the strands and a molecular barcode towards another end. The primer hybridization site can be a sequencing initiation sequence for subsequent sequencing. The library adaptor is less than or equal to 80 nucleotide bases in length. In **FIG. 9B**, the library adaptors are ligated at both ends of the target polynucleotide molecule to provide a tagged target polynucleotide molecule. The tagged target

polynucleotide molecule may be subjected to nucleic acid amplification to generate copies of the target. Next, in **FIG. 9C**, sequencing adaptors containing sequencer motifs are provided and hybridized to the tagged target polynucleotide molecule. The sequencing adaptors contain sample identification motifs. The sequencing adaptors can contain sequences to permit sequencing of the tagged target with a given sequencer.

[00152] **D. Sequencing**

[00153] Tagged polynucleotides can be sequenced to generate sequence reads (*e.g.*, as shown in step (106), **FIG. 1**). For example, a tagged duplex polynucleotide can be sequenced. Sequence reads can be generated from only one strand of a tagged duplex polynucleotide. Alternatively, both strands of a tagged duplex polynucleotide can generate sequence reads. The two strands of the tagged duplex polynucleotide can comprise the same tags. Alternatively, the two strands of the tagged duplex polynucleotide can comprise different tags. When the two strands of the tagged duplex polynucleotide are differently tagged, sequence reads generated from one strand (*e.g.*, a Watson strand) can be distinguished from sequence reads generated from the other strands (*e.g.*, a Crick strand). Sequencing can involve generating multiple sequence reads for each molecule. This occurs, for example, as a result the amplification of individual polynucleotide strands during the sequencing process, *e.g.*, by PCR.

[00154] Methods disclosed herein can comprise amplifying of polynucleotides. Polynucleotides amplification can result in the incorporation of nucleotides into a nucleic acid molecule or primer thereby forming a new nucleic acid molecule complementary to a template nucleic acid. The newly formed polynucleotide molecule and its template can be used as templates to synthesize additional polynucleotides. The polynucleotides being amplified can be any nucleic acids, for example, deoxyribonucleic acids, including genomic DNAs, cDNAs (complementary DNA), cfDNAs, and circulating tumor DNAs (ctDNAs). The polynucleotides being amplified can also be RNAs. As used herein, one amplification reaction may comprise many rounds of DNA replication. DNA amplification reactions can include, for example, polymerase chain reaction (PCR). One PCR reaction may comprise 2-100 "cycles" of denaturation, annealing, and synthesis of a DNA molecule. For example, 2-7, 5-10, 6-11, 7-12, 8-13, 9-14, 10-15, 11-16, 12-17, 13-18, 14-19, or 15-20 cycles can be performed during the amplification step. The condition of the PCR can be optimized based on the GC content of the sequences, including the primers.

[00155] Nucleic acid amplification techniques can be used with the assays described herein. Some amplification techniques are the PCR methodologies which can include, but are not limited to, solution PCR and *in situ* PCR. For example, amplification may comprise PCR-based amplification. Alternatively, amplification may comprise non PCR-based amplification. Amplification of the template nucleic acid may comprise use of one or more polymerases. For example, the polymerase may be a DNA polymerase or an RNA polymerase. In some cases, high fidelity amplification is performed such as with the use of high fidelity polymerase (e.g., Phusion® High-Fidelity DNA Polymerase) or PCR protocols. In some cases, the polymerase may be a high fidelity polymerase. For example, the polymerase may be KAPA HiFi DNA polymerase. The polymerase may also be Phusion DNA polymerase. The polymerase may be used under reaction conditions that reduce or minimize amplification biases, e.g., due to fragment length, GC content, etc.

[00156] Amplification of a single strand of a polynucleotide by PCR will generate copies both of that strand and its complement. During sequencing, both the strand and its complement will generate sequence reads. However, sequence reads generated from the complement of, for example, the Watson strand, can be identified as such because they bear the complement of the portion of the duplex tag that tagged the original Watson strand. In contrast, a sequence read generated from a Crick strand or its amplification product will bear the portion of the duplex tag that tagged the original Crick strand. In this way, a sequence read generated from an amplified product of a complement of the Watson strand can be distinguished from a complement sequence read generated from an amplification product of the Crick strand of the original molecule.

[00157] All amplified polynucleotides can be submitted to a sequencing device for sequencing. Alternatively, a sampling, or subset, of all of the amplified polynucleotides is submitted to a sequencing device for sequencing. With respect to any original double-stranded polynucleotide there can be three results with respect to sequencing. First, sequence reads can be generated from both complementary strands of the original molecule (that is, from both the Watson strand and from the Crick strand). Second, sequence reads can be generated from only one of the two complementary strands (that is, either from the Watson strand or from the Crick strand, but not both). Third, no sequence read may be generated from either of the two complementary strands. Consequently, counting unique sequence reads mapping to a genetic locus will underestimate the number of double-stranded

polynucleotides in the original sample mapping to the locus. Described herein are methods of estimating the unseen and uncounted polynucleotides.

[00158] The sequencing method can be massively parallel sequencing, that is, simultaneously (or in rapid succession) sequencing any of at least 100, 1000, 10,000, 100,000, 1 million, 10 million, 100 million, or 1 billion polynucleotide molecules. 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, Maxam-Gilbert or Sanger sequencing, primer walking, sequencing using PacBio, SOLiD, Ion Torrent, or Nanopore platforms and any other sequencing methods known in the art.

[00159] For example, duplex-tagged polynucleotides can be amplified, by for example PCR (*see e.g.*, **FIG. 4A** duplex-tagged polynucleotides are referred to as mm' and nn'). In **Fig. 4A**, the strand of the duplex polynucleotide including sequence m bears sequence tags w and y, while the strand of the duplex polynucleotide including sequence m' bears sequence tags x and z. Similarly, the strand of the duplex polynucleotide including sequence n bears sequence tags a and c, while the strand of the duplex polynucleotide including sequence n' bears sequence tags b and d. During amplification, each strand produces itself and its complementary sequence. However, for example, an amplification progeny of original strand m that includes the complementary sequence, m', is distinguishable from an amplification progeny of original strand m' because the progeny from original strand m will have the sequence 5'-y'm'w'-3' and the progeny of the original m' strand one strand will have the sequence 5'-zm'x-3'. **FIG. 4B** shows amplification in more detail. During amplification, errors can be introduced into the amplification progeny, represented by dots. The application progeny are sampled for sequencing, so that not all strands produce sequence reads, resulting in the sequence reads indicated. Because sequence reads can come from either of a strand or its complement, both sequences and complement sequences will be included in the set of sequence reads. It should be noted that it is possible that a polynucleotide would bear the same tag on each end. Thus, for a tag "a", and polynucleotide "m", a first strand could be tagged a-m-a', and the complement could be tagged a-m'-a.

[00160] **E. Determining consensus sequence reads**

[00161] Methods disclosed herein can comprise determining consensus sequence reads in sequence reads (*e.g.*, as shown in step (108), **FIG. 1**), such as by reducing or tracking redundancy. Sequencing of amplified polynucleotides can produce reads of the several amplification products from the same original polynucleotide, referred to as "redundant reads". By identifying redundant reads, unique molecules in the original sample can be determined. If the molecules in a sample are uniquely tagged, then reads generated from amplification of a single unique original molecule can be identified based on their distinct barcode. Ignoring barcodes, reads from unique original molecules can be determined based on sequences at the beginning and end of a read, optionally in combination with the length of the read. In certain cases, however, a sample may be expected to have a plurality of original molecules having the same start stop sequences and the same length. Without barcoding, these molecules are difficult to distinguish from one another. However, if a collection of polynucleotides is non-uniquely tagged (that is, an original molecule shares the same identifier with at least one other original molecule), combining information from a barcode with start/stop sequence and/or polynucleotide length significantly increases the probability that any sequence read can be traced back to an original polynucleotide. This is because, in part, even without unique tagging, it is unlikely that any two original polynucleotides having the same start/stop sequence and length also will be tagged with the same identifier.

[00162] **F. Collapsing**

[00163] Collapsing allows for reduction in noise (*i.e.*, background) that is generated at each step of the process. Methods disclosed herein can comprise collapsing, *e.g.*, generating a consensus sequence by comparing multiple sequence reads. For example, sequence reads generated from a single original polynucleotide can be used to generate a consensus sequence of that original polynucleotide. Iterative rounds of amplification can introduce errors into progeny polynucleotides. Also, sequencing typically may not be performed with perfect fidelity so sequencing errors are introduced at this stage as well. However, comparison of sequence reads of molecules derived from a single original molecule, including those that have sequence variants, can be analyzed so as to determine the original, or "consensus" sequence. This can be done phylogenetically. Consensus sequences can be generated from families of sequence reads by any of a variety of methods. Such methods include, for example, linear or non-linear methods of building consensus sequences (such as voting (*e.g.*, biased 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. For example, if all or most of the sequence reads tracking back to an original molecule bear the same sequence variant, that variant probably existed in the original molecule. On the other hand, if a sequence variant exists in a subset of redundant sequence reads, that variant may have been introduced during amplification/sequencing and represents an artifact not existing in the original. Furthermore, if only sequence reads derived from the Watson or Crick strand of an original polynucleotide contain the variant, the variant may have been introduced through single-sided DNA damage, first-cycle PCR error or through contaminating polynucleotides that were amplified from a different sample.

[00164] After fragments are amplified and the sequences of amplified fragments are read and aligned, the fragments are subjected to base calling, *e.g.*, determining for each locus the most likely nucleotide. However, variations in the number of amplified fragments and unseen amplified fragments (*e.g.*, those without being read their sequences; reasons could be too many such as amplification errors, sequencing reading errors, too long, too short, being chopped, etc.) may introduce errors in base calling. If there are too many unseen amplified fragments with respect to the seen amplified fragments (amplified fragments actually being read), the reliability of base calling may be diminished.

[00165] Therefore, disclosed herein is a method to correct for the number of unseen fragments in base calling. For example, when base calling for locus A (an arbitrary locus), it is first assumed that there are N amplified fragments. The sequence readouts can come from two types of fragments: double-strand fragments and single-strand fragments. Therefore, we assign N1, N2, and N3 as the numbers of double-strands, single-strands, and unseen fragments, respectively. Thus, N=N1+N2+N3 (N1 and N2 are known from the sequence readouts, and N and N3 are unknown). If the formula is solved for N (or N3), then N3 (or N) will be inferred.

[00166] Probability is used to estimate N. For example, we assign "p" to be the probability of having detected (or having read) a nucleotide of locus A in a sequence readout of a single-strand.

[00167] For sequence readouts from double-strands, the nucleotide call from a double-strand amplified fragment has a probability of p * p=p^2, seeing all N1 double-strands has the following equation: N1=N * (p^2).

[00168] For sequence readouts from a single-strand. Assuming that one of the 2 strands is seen, and the other is unseen, the probability of seeing one strand is "p", but the probability of missing the other strand is (1-p). Furthermore, by not distinguishing the single strand sourcing from 5-primer and sourcing from 3-primer, there is a factor of 2. Therefore, the nucleotide call from a single-strand amplified fragment has a probability 2×p×(1-p). Thus, seeing all N2 single-strands has the following equation: N2=N×2×p×(1-p).

[00169] "p" is also unknown. To solve p, the ratio of N1 to N2 is used to solve for "p":

$$R = \frac{N1}{N2} = \frac{Np^2}{2Np(1-p)} = \frac{p^2}{2p(1-p)} = \frac{p}{2(1-p)}$$

Once "p" is found, N can be found. After N is found, can be found N3 = N − N1 − N2.

[00170] Besides the ratio of paired versus unpaired strands (which is a measure post-collapsing), there is useful information in the pre-collapsing read depth at each locus. This information can be used to further improve the call for total molecule count and/or increase confidence of calling variants.

[00171] For example, **FIG. 4C** demonstrates sequence reads corrected for complementary sequences. Sequences generated from an original Watson strand or an original Crick strand can be differentiated on the basis of their duplex tags. Sequences generated from the same original strand can be grouped. Examination of the sequences can allow one to infer the sequence of the original strand (the "consensus sequence"). In this case, for example, the sequence variant in the nn' molecule is included in the consensus sequence because it included in every sequence read while other variants are seen to be stray errors. After collapsing sequences, original polynucleotide pairs can be identified based on their complementary sequences and duplex tags.

[00172] **FIG. 5** demonstrates increased confidence in detecting sequence variants by pairing reads from Watson and Crick strands. Sequence nn' can include a sequence variant indicated by a dot. In some cases, sequence pp' does not include a sequence variant. Amplification, sequencing, redundancy reduction and pairing can result in both Watson and Crick strands of the same original molecule including the sequence variant. In contrast, as a result of errors introduced during amplification and sampling during sequencing, the consensus sequence of the Watson strand p can contain a sequence variant, while the consensus sequence of the Crick strand p' does not. It is less likely that amplification and sequencing will introduce the same variant into both strands (nn' sequence) of a duplex than onto one strand (pp' sequence). Therefore, the variant in the pp' sequence is more likely to

be an artifact, and the variant in the nn' sequence is more likely to exist in the original molecule.

[00173] Methods disclosed herein can be used to correct errors resulted from experiments, *e.g.*, PCR, amplification, and/or sequencing. For example, such a method can comprises attaching one or more double stranded adaptors to both ends of a double stranded polynucleotide, thereby providing a tagged double stranded polynucleotide; amplifying the double stranded tagged polynucleotide; sequencing both strands of the tagged polynucleotide; comparing the sequence of one strand with its complement to determine any errors introduced during sequencing; and correcting errors in the sequence based on (d). The adaptors used in this method can be any adaptors disclosed herein, *e.g.*, Y-shaped adaptors. The adaptor can comprise any barcodes (*e.g.*, distinct barcodes) disclosed herein.

[00174] **G. Mapping**

[00175] Sequence reads or consensus sequences can be mapped to one or more selected genetic loci (*e.g.*, as shown step (110), **FIG. 1**). A genetic locus can be, for example, a specific nucleotide position in the genome, a sequence of nucleotides (for example, an open reading frame), a fragment of a chromosome, a whole chromosome, or an entire genome. A genetic locus can be a polymorphic locus. Polymorphic locus can be a locus at which sequence variation exists in the population and/or exists in a subject and/or a sample. A polymorphic locus can be generated by two or more distinct sequences coexisting at the same location of the genome. The distinct sequences can differ from one another by one or more nucleotide substitutions, a deletion/insertion, and/or a duplication of any number of nucleotides, generally a relatively small number of nucleotides, such as less than 50, 45, 40, 35, 30, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, or 1 nucleotide(s), among others. A polymorphic locus can be created by a single nucleotide position that varies within the population, *e.g.* a single nucleotide variation (SNV) or a single nucleotide polymorphism (SNP).

[00176] A reference genome for mapping can include the genome of any species of interest. Human genome sequences useful as references can include the hg19 assembly or any previous or available hg assembly. Such sequences can be interrogated using the genome browser available at genome.ucsc.edu/index.html. Other species genomes include, for example PanTro2 (chimp) and mm9 (mouse).

[00177] In methods disclosed herein, collapsing can be performed before or after mapping. In some aspects, collapsing can be performed before mapping. For example, sequence reads

can be grouped into families based on their tags and one or more endogenous sequences, without regard to where the reads map in the genome. Then, the members of a family can be collapsed into a consensus sequence. The consensus sequence can be generated using any collapsing method disclosed herein. Then the consensus sequence can be mapped to locations in the genome. Reads mapped to a locus can be quantified (*e.g.*, counted). Percentage of reads carrying a mutation at a locus can also be determined. Alternatively, collapsing can be performed after mapping. For example, all reads can first be mapped to the genome. Then the reads can be grouped into families based on their tags and one or more endogenous sequences. Since the reads have been mapped to the genome, consensus bases can be determined for each family at each locus. In other aspects, consensus sequence can be generated for one strand of a DNA molecule (*e.g.*, for a Watson strand or a Crick strand). Mapping can be performed before or after the consensus sequence for one strand of the DNA molecule is determined. Numbers of Doublets and Singlets can be determined. These numbers can be used to calculate unseen molecules. For example, the unseen molecules can be calculated using the following equation: N= D+S+U; D=Np(2), S=N2pq, where p=1-q, where p is the probability of seeing; q is the probability of missing a strand.

[00178] **H. <u>Grouping</u>**

[00179] Methods disclosed herein can also comprise grouping sequence reads. Sequence reads can be grouped based on various types of sequences, *e.g.*, sequences of an oligonucleotide tag (*e.g.*, a barcode), sequence of a polynucleotide fragments, or any combinations. For example, as shown in step (112) (**FIG. 1**), sequence reads can be grouped as follows: Sequence reads generated from a "Watson" strand and those generated from a "Crick" strand of a double-stranded polynucleotide in the sample are identifiable based on the duplex tags that they bear. In this way, a sequence read or consensus sequence from a Watson strand of a duplex polynucleotide can be paired with a sequence read or consensus sequence from its complementary Crick strand. Paired sequence reads are referred to as a "Pair".

[00180] Sequence reads for which no sequence read corresponding to a complementary strand can be found among the sequence reads are termed "Singlets".

[00181] Double-stranded polynucleotides for which a sequence read for neither of the two complementary strands has been generated are referred to as "Unseen" molecules.

[00182] **I. Quantifying**

[00183] Methods disclosed herein also comprise quantifying sequence reads. For example, as shown in step (114) (**FIG. 1**), Pairs and Singlets mapping to a selected genetic locus, or to each of a plurality of selected genetic loci, are quantified, *e.g.*, counted.

[00184] The quantifying can comprise estimating number of polynucleotides in the sample (*e.g.*, Pairs polynucleotides, Singlets polynucleotides, or Unseen polynucleotides. For example, as shown in step (116) (**FIG. 1**), the number of double-stranded polynucleotides in the sample for which no sequence reads were generated ("Unseen" polynucleotides) is estimated. The probability that a double strand polynucleotide generates no sequence reads can be determined based on the relative number of Pairs and Singlets at any locus. Using this probability, the number of Unseen polynucleotide can be estimated.

[00185] In step (118) an estimate for the total number of double-stranded polynucleotides in a sample mapping to a selected locus is the sum of the number of Pairs, the number of Singlets and the number of Unseen molecules mapping to the locus.

[00186] The number of Unseen original molecules in a sample can be estimated based on the relative number of Pairs and Singlets (**FIG. 2**). Referring to **FIG. 2**, as an example, counts for a particular genomic locus, Locus A, are recorded, where 1000 molecules are paired and 1000 molecules are unpaired. Assuming a uniform probability, p, for an individual Watson or Crick strand to make it through the process subsequent to conversion, one can calculate the proportion of molecules that fail to make it through the process (Unseen) as follows: Let R = ratio of paired to unpaired molecules = 1, so $R=1=p^2/(2p(1-p))$. This implies that p=2/3 and that the quantity of lost molecules is equal to $(1-p)^2 = 1/9$. Thus in this example, approximately 11% of converted molecules are lost and never detected. Consider another genomic locus, Locus B, in the same sample where 1440 molecules are paired and 720 are unpaired. Using the same method, we can infer the number of molecules that are lost, is only 4%. Comparing the two areas, it may be assumed that Locus A had 2000 unique molecules as compared to 2160 molecules in Locus B – a difference of almost 8%. However, by correctly adding in the lost molecules in each region, we infer there are 2000/(8/9)=2250 molecules in Locus A and 2160/.96=2250 molecules in Locus B. Hence, the counts in both regions are actually equal. This correction and thus much higher sensitivity can be achievable by converting the original double-stranded nucleic acid molecules and bioinformatically keeping track of all those that are paired and unpaired at the end of the process. Similarly, the same procedure can be used to infer true copy number

variations in regions that appear to have similar counts of observed unique molecules. By taking the number of unseen molecules into consideration in the two or more regions, the copy number variation becomes apparent.

[00187] In addition to using binomial distribution, other methods of estimating numbers of unseen molecules include exponential, beta, gamma or empirical distributions based on the redundancy of sequence reads observed. In the latter case, the distribution of read counts for paired and unpaired molecules can be derived from such redundancy to infer the underlying distribution of original polynucleotide molecules at a particular locus. This can often lead to a better estimation of the number of unseen molecules.

[00188] **J. CNV Detection**

[00189] Methods disclosed herein also comprise detecting CNV. For example, as shown in step (120) (**FIG. 1**), once the total number of polynucleotides mapping to a locus is determined, this number can be used in standard methods of determining CNV at the locus. A quantitative measure can be normalized against a standard. The standard can be an amount of any polynucleotides. In one method, a quantitative measure at a test locus can be standardized against a quantitative measure of polynucleotides mapping to a control locus in the genome, such as gene of known copy number. Quantitative measures can be compared against the amount of nucleic acid in any sample disclosed herein. For example, in another method, the quantitative measure can be compared against the amount of nucleic acid in the original sample. For example, if the original sample contained 10,000 haploid gene equivalents, the quantitative measure can be compared against an expected measure for diploidy. In another method, the quantitative measure can be normalized against a measure from a control sample, and normalized measures at different loci can be compared.

[00190] 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.

[00191] After the sequence read coverage ratios have been determined, a stochastic modeling algorithm can be optionally 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.

[00192] Methods disclosed herein can comprise detecting SNVs, CNVs, insertions, deletions, and/or rearrangements at a specific region in a genome. The specific genomic region can comprise a sequence in a gene, such as ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, or NTRK1.

[00193] In some cases, the method uses a panel which comprises exons of one or more genes. The panel can comprise introns of one or more genes as well. The panel can also comprise exons and introns of one or more genes. The one or more genes can be those disclosed above. The panel can comprise about 80,000 bases which cover a panel of genes. The panel can comprise about 1000, 2000, 3000, 4000, 5000, 10000, 15000, 20000, 25000, 30000, 35000, 40000, 45000, 50000, 55000, 60000, 65000, 70000, 75000, 80000, 85000, 90000, 95000, 100000, 105000, 110000, 115000, 120000, 125000, or more bases.

[00194] In some aspects, copy number of a gene can be reflected in the frequency of a genetic form of the gene in a sample. For example, in a healthy individual, no copy number variation is reflected in a variant in a gene in one chromosome (*e.g.*, heterozygosity) being detected in about 50% of detected molecules in a sample. Also, in a healthy individual, duplication of a gene bearing a variant can be reflected in the variant being detected in about 66% of detected molecules in a sample. Accordingly, if the tumor burden in a DNA sample is 10%, the frequency of a somatic mutation in a gene in one chromosome of cancer cells, without CNV, can be about 5%. The converse can be true in the case of aneuploidy.

[00195] The methods disclosed herein can be used to determine whether a sequence variant is more likely present in the germ line level or resulted from a somatic cell mutation, *e.g.*, in a cancer cell. For example, a sequence variant in a gene detected at levels arguably consistent with heterozygosity in the germ line is more likely the product of a somatic mutation if CNV is also detected in that gene. In some cases, to the extent we expect that a gene duplication in the germ line bears a variant consistent with genetic dose (*e.g.*, 66% for trisomy at a locus), detection gene amplification with a sequence variant dose that deviates significantly from this expected amount indicates that the CNV is more likely present as a result of somatic cell mutation.

[00196] The methods disclosed herein can also be used to infer tumor heterogeneity in a situation in which sequence variants in two genes are detected at different frequencies. For example, tumor heterogeneity can be inferred when two genes are detected at different frequencies but their copy numbers are relatively equal. Alternatively, tumor homogeneity can be inferred when the difference in frequency between two sequence variants is consistent with difference in copy number for the two genes. Thus, for example, if an EGFR variant is detected at 11% and a KRAS variant is detected at 5%, and no CNV is detected at these genes, the difference in frequency likely reflects tumor heterogeneity (*e.g.*, all tumor cells carry an EGFR mutant and half the tumor cells also carry a KRAS mutant). Alternatively, if the EGFR gene carrying the mutant is detected at 2-times normal copy number, one interpretation is a homogenous population of tumor cells, each cell carrying a mutant in the EGFR and KRAS genes, but in which the KRAS gene is duplicated.

[00197] In response to chemotherapy, a dominant tumor form can eventually give way through Darwinian selection to cancer cells carrying mutants that render the cancer unresponsive to the therapy regimen. Appearance of these resistance mutants can be delayed through methods of this invention. In one embodiment of this method, a subject is subjected to one or more pulsed therapy cycles, each pulsed therapy cycle comprising a first period during which a drug is administered at a first amount and a second cycle during which the drug is administered at a second, reduced amount. The first period can be characterized by a tumor burden detected above a first clinical level. The second period can be characterized by a tumor burden detected below a second clinical level. First and second clinical levels can be different in different pulsed therapy cycles. For example, the first clinical level can be lower in succeeding cycles. A plurality of cycles can include at least 2, 3, 4, 5, 6, 7, 8 or more cycles. For example, the BRAF mutant V600E may be detected in polynucleotides of a

disease cell at an amount indicating a tumor burden of 5% in cfDNA. Chemotherapy can commence with dabrafenib. Subsequent testing can show that the amount of the BRAF mutant in the cfDNA falls below 0.5% or to undetectable levels. At this point, dabrafenib therapy can stop or be significantly curtailed. Further subsequent testing may find that DNA bearing the BRAF mutation has risen to 2.5% of polynucleotides in cfDNA. At this point, dabrafenib therapy can be re-started, *e.g.*, at the same level as the initial treatment. Subsequent testing may find that DNA bearing the BRAF mutation has decreased to 0.5% of polynucleotides in cfDNA. Again, dabrafenib therapy can be stopped or reduced. The cycle can be repeated a number of times.

[00198] A therapeutic intervention can also be changed upon detection of the rise of a mutant form resistant to an original drug. For example, cancers with the EGFR mutation L858R respond to therapy with erlotinib. However, cancers with the EGFR mutation T790M are resistant to erlotinib. However, they are responsive to ruxolitinib. A method of this invention involves monitoring changes in tumor profile and changing a therapeutic intervention when a genetic variant associated with drug resistance rises to a predetermined clinical level.

[00199] Methods disclosed in this invention can comprise a method of detecting disease cell heterogeneity from a sample comprising polynucleotides from somatic cells and disease cells, the method comprising: a) quantifying polynucleotides in the sample bearing a sequence variant at each of a plurality of genetic loci; b) determining CNV at each of the plurality of genetic loci; different relative amounts of disease molecules at a locus, wherein the CNV indicates a genetic dose of a locus in the disease cell polynucleotides; c) determining a relative measure of quantity of polynucleotides bearing a sequence variant at a locus per genetic dose at the locus for each of a plurality of the loci; and d) comparing the relative measures at each of the plurality of loci, wherein different relative measures indicates tumor heterogeneity. In the methods disclosed herein, the genetic dose can be determined on a total molecule basis. For example, if there are 1X total molecules at a first locus, and 1.2X molecules mapped to a second locus, then the genetic dose is 1.2. Variants at this locus can be divided by 1.2. In some aspects, the method disclosed herein can be used to detect any disease cell heterogeneity, *e.g.*, tumor cell heterogeneity. The methods can be used to detect disease cell heterogeneity from a sample comprising any types of polynucleotides, *e.g.*, cfDNA, genomic DNA, cDNA, or ctDNA. In the methods, the quantifying can comprise, for example, determining the number or relative amount of the polynucleotides. Determining

CNV can comprise mapping and normalizing different relative amounts of total molecules to a locus.

[00200] In another aspect, in response to chemotherapy, a dominant tumor form can eventually give way through Darwinian selection to cancer cells carrying mutants that render the cancer unresponsive to the therapy regimen. Appearance of these resistance mutants can be delayed through methods disclosed throughout. The methods disclosed herein can comprise a method comprising: a) subjecting a subject to one or more pulsed therapy cycles, each pulsed therapy cycle comprising (i) a first period during which a drug is administered at a first amount and (ii) a second period during which the drug is administered at a second, reduced amount; wherein (A) the first period is characterized by a tumor burden detected above a first clinical level; and (B) the second period is characterized by a tumor burden detected below a second clinical level.

[00201] **K. <u>Sequence Variant Detection</u>**

[00202] Systems and methods disclosed herein can be used to detect sequence variants, *e.g.*, SNVs. For example, a sequence variant can be detected from consensus sequences from multiple sequence reads, for example, from at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 21, at least 22, at least 23, at least 24, at least 25, at least 26, at least 27, at least 28, at least 29, at least 30, at least 31, at least 32, at least 33, at least 34, at least 35, at least 36, at least 37, at least 38, at least 39, at least 40, at least 41, at least 42, at least 43, at least 44, at least 45, at least 46, at least 47, at least 48, at least 49, at least 50, at least 51, at least 52, at least 53, at least 54, at least 55, at least 56, at least 57, at least 58, at least 59, at least 60, at least 61, at least 62, at least 63, at least 64, at least 65, at least 66, at least 67, at least 68, at least 69, at least 70, at least 71, at least 72, at least 73, at least 74, at least 75, at least 76, at least 77, at least 78, at least 79, at least 80, at least 81, at least 82, at least 83, at least 84, at least 85, at least 86, at least 87, at least 88, at least 89, at least 90, at least 91, at least 92, at least 93, at least 94, at least 95, at least 96, at least 97, at least 98, at least 99, at least 100, at least 200, at least 300, at least 400, at least 500, at least 600, at least 700, at least 800, at least 900, at least 1000, at least 2000, at least 3000, at least 4000, at least 5000, at least 6000, at least 7000, at least 8000, at least 9000, at least 10000 or more sequence reads. A consensus sequence can be from sequence reads of a single strand polynucleotide. A consensus sequence can also be from sequence reads of one strand of a double-stranded polynucleotide (*e.g.*, pairing reads).

In an exemplary method, pairing reads allows one to identify with increased confidence the existence of a sequence variant in a molecule. For example, if both strands of a Pair include the same variant, one can be reasonably sure that the variant existed in the original molecule, as the chance that the same variant is introduced into both strands during amplification/sequencing is rare. In contrast, if only one strand of a Pair includes the sequence variant, this is more likely to be an artifact. Similarly, the confidence that a Singlet bearing a sequence variant existed in the original molecule is less than the confidence if the variant exists in a Duplex, as there is higher probability that the variant can be introduced once than twice during amplification/sequencing.

[00203] Other methods of copy number variation detection and the sequence variant detection are described in PCT/US2013/058061, which is entirely incorporated herein by reference.

[00204] Sequence reads can be collapsed to generate a consensus sequence, which can be mapped to a reference sequence to identify genetic variants, such as CNV or SNV. As an alternative, the sequence reads are mapped prior to or even without mapping. In such a case, the sequence reads can be individually mapped to the reference to identify a CNV or SNV.

[00205] **FIG. 3** shows a reference sequence encoding a genetic Locus A. The polynucleotides in FIG. 3 may be Y-shaped or have other shapes, such as hairpin.

[00206] In some cases, an SNV or multiple-nucleotide variant (MNV) can be determined across multiple sequence reads at a given locus (e.g., nucleotide base) by aligning sequence reads that correspond to that locus. Next, a plurality of sequential nucleotide bases from at least a subset of the sequence reads are mapped to the reference to a SNV or MNV in a polynucleotide molecule or portion thereof that corresponds to the reads. The plurality of sequential nucleotide bases can span an actual, inferred or suspected location of the SNV or MNV. The plurality of sequential nucleotide bases can span at least 3, 4, 5, 6, 7, 8, 9, or 10 nucleotide bases.

[00207] **L. Detecting/Quantifying Nucleic Acids**

[00208] The methods described throughout can be used to tag nucleic acids fragments, such as deoxyribonucleic acid (DNA), at extremely high efficiency. This efficient tagging allows a person to efficiently and accurately detect rare DNA in heterogenous populations of original DNA fragments (such as in cfDNA). A rare polynucleotide (e.g., rare DNA) can be a polynucleotide that comprises a genetic variant occurring in a population of polynucleotides at a frequency of less than 10%, 5%, 4%, 3%, 2%, 1%, or 0.1%. A rare DNA can be a

polynucleotide with a detectable property at a concentration less than 50%, 25%, 10%, 5%, 1%, or 0.1%

[00209] Tagging can occur in a single reaction. In some cases, two or more reactions can be performed and pooled together. Tagging each original DNA fragments in a single reaction can result in tagging such that greater than 50% (*e.g.*, 60%, 70%, 80%, 90%, 95%, or 99%) of the original DNA fragments are tagged at both ends with tags that comprise molecular barcodes, thereby providing tagged DNA fragments. Tagging can also result in greater than 30%, 35%, 40%, 45%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% of the original DNA fragments tagged at both ends with tags that comprise molecular barcodes. Tagging can also result in 100% of the original DNA fragments tagged at both ends with tags that comprise molecular barcodes. Tagging can also result in single end tagging.

[00210] Tagging can also occur by using an excess amount of tags as compared to the original DNA fragments. For example, the excess can be at least 5-fold excess. In other cases, the excess can be at least 1.25, 1.5, 1.75, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100 or more fold excess. Tagging can comprise attachment to blunt ends or sticky ends. Tagging can also be performed by hybridization PCR. Tagging can also be performed in low reaction volumes, such as 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, or 100 pico- and/or microliters.

[00211] The method can also include performing high fidelity amplification on the tagged DNA fragments. Any high fidelity DNA polymerases can be used. For example, the polymerase may be KAPA HiFi DNA polymerase or Phusion DNA polymerase.

[00212] Further, the method can comprise selectively enriching a subset of the tagged DNA fragments. For example, selective enrichment can be performed by hybridization or amplification techniques. The selective enrichment can be performed using a solid support (*e.g.*, beads). The solid support (*e.g.*, beads) can comprise probes (*e.g.*, oligonucleotides specifically hybridizing to certain sequences. For example, the probes can hybridize with

certain genomic regions, *e.g.*, genes. In some cases, the genomic regions, *e.g.*, genes, can be regions associated with diseases, *e.g.*, cancer. After enrichment, the selected fragmented can be attached any sequencing adaptor disclosed in this invention. For example, a sequence adaptor can comprise a flow cell sequence, a sample barcode, or both. In another example, a sequence adaptor can be a hairpin shaped adaptor and/or comprises a sample barcode. Further, the resulting fragments can be amplified and sequenced. In some cases, the adaptor does not comprise a sequencing primer region.

[00213] The method can include sequencing one or both strands of the DNA fragments. In one case, both strands of the DNA fragment are independently sequenced. The tagged, amplified, and/or selectively enriched DNA fragments are sequenced to obtain sequence reads that comprise sequence information of the molecular barcodes and at least a portion of the original DNA fragments.

[00214] The method can include reducing or tracking redundancy (as described above) in the sequence reads to determine consensus reads that are representative of single-strands of the original DNA fragments. For example, to reduce or track redundancy, the method can include comparing sequence reads having the same or similar molecular barcodes and the same or similar end of fragment sequences. The method can comprise performing a phylogentic analysis on the sequence reads having the same or similar molecular barcodes. The molecular barcodes can have a barcode with varying edit distances (including any edit distances as described throughout), for example, an edit distance of up to 3. The end of the fragment sequences can include fragment sequences having an edit distance with varying distances (including any edit distances as described throughout), for example, an edit distance of up to 3.

[00215] The method can comprise binning the sequence reads according to the molecular barcodes and sequence information. For example, binning the sequence reads according to the molecular barcodes and sequence information can be performed from at least one end of each of the original DNA fragments to create bins of single stranded reads. The method can further comprise in each bin, determining a sequence of a given original DNA fragment among the original DNA fragments by analyzing sequence reads.

[00216] In some cases, sequence reads in each bin can be collapsed to a consensus sequence and subsequently mapped to a genome. As an alternative, sequence reads can be mapped to a genome prior to binning and subsequently collapsed to a consensus sequence.

[00217] The method can also comprise sorting sequence reads into paired reads and unpaired reads. After sorting, the number of paired reads and unpaired reads that map to each of one or more genetic loci can be quantified.

[00218] The method can include quantifying the consensus reads to detect and/or quantify the rare DNA, which are described throughout. The method can comprise detecting and/or quantifying the rare DNA by comparing a number of times each base occurs at each position of a genome represented by the tagged, amplified, and/or enriched DNA fragments.

[00219] The method can comprise tagging the original DNA fragments in a single reaction using a library of tags. The library can include at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 50, at least 100, at least 500, at least 1000, at least 5000, at least 10000, or any number of tags as disclosed throughout. For example, the library of tags can include at least 8 tags. The library of tags can include 8 tags (which can generate 64 different possible combinations). The method can be conducted such that a high percentage of fragments, *e.g.*, greater than 50% (or any percentages as described throughout) are tagged at both ends, wherein each of the tags comprises a molecular barcode.

[00220] M. Processing and/or Analyzing Nucleic Acids

[00221] The methods described throughout can be used for processing and/or analyzing a nucleic acid sample of a subject. The method can comprising exposing polynucleotide fragments of the nucleic acid sample to a plurality of polynucleotide molecules to yield tagged polynucleotide fragments. The plurality of polynucleotide molecules that can be used are described throughout the application.

[00222] For example, the plurality of polynucleotide molecules can be each less than or equal to 40 nucleic acid bases in length and have distinct barcode sequences with respect to at least 4 nucleic acid bases and an edit distance of at least 1, wherein each of the distinct barcode sequences is within 20 nucleic acid bases from a terminal end of a respective one of the plurality of polynucleotide molecules, and wherein the plurality of polynucleotide molecules are not sequencing adaptors.

[00223] The tagged polynucleotide fragments can be subjected to nucleic acid amplification reactions under conditions that yield amplified polynucleotide fragments as amplification products of the tagged polynucleotide fragments. After amplification, the nucleotide sequence of the amplified tagged polynucleotide fragments is determined. In

some cases, the nucleotide sequences of the amplified tagged polynucleotide fragments are determined without the use of polymerase chain reaction (PCR).

[00224] The method can comprise analyzing the nucleotide sequences with a programmed computer processor to identify one or more genetic variants in the nucleotide sample of the subject. Any genetic alterations can be identified, including but not limited to, base change(s), insertion(s), repeat(s), deletion(s), copy number variation(s) , epigenetic modification(s), nucleosome binding site(s), copy number change(s) due to origin(s) of replication, and transversion(s). Other genetic alterations can include, but are not limited to, one or more tumor associated genetic alterations.

[00225] The subject of the methods can be suspected of having a disease. For example, the subject can be suspected of having cancer. The method can comprise collecting a nucleic acid sample from a subject. The nucleic acid sample can be collected from blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool, cerebral spinal fluid, skin, hair, sweat, and/or tears. The nucleic acid sample can be a cell-free nucleic acid sample. In some cases, the nucleic acid sample is collected from no more than 100 nanograms (ng) of double-stranded polynucleotide molecules of the subject.

[00226] The polynucleotide fragments can comprise double-stranded polynucleotide molecules. In some cases, the plurality of polynucleotide molecules are coupled to the polynucleotide fragments via blunt end ligation, sticky end ligation, molecular inversion probes, polymerase chain reaction (PCR), ligation-based PCR, multiplex PCR, single strand ligation, or single strand circularization.

[00227] The method as described herein results in high efficiency tagging of nucleic acids. For example, exposing the polynucleotide fragments of the nucleic acid sample to the plurality of polynucleotide molecules yields the tagged polynucleotide fragments with a conversion efficiency of at least 30%, *e.g.*, of at least 50% (*e.g.*, 60%, 70%, 80%, 90%, 95%, or 99%). Conversion efficiency of at least 30%, 35%, 40%, 45%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% can be achieved.

[00228] The method can result in a tagged polynucleotide fragment that share common polynucleotide molecules. For example, any of at least 5%, 6%, 7%, 8%, 9%, 10%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%,

97%, 98%, 99% or 100% of the tagged polynucleotide fragments share a common polynucleotide molecule. The method can comprise generating the polynucleotide fragments from the nucleic acid sample.

[00229] In some cases, the subjecting of the method comprises amplifying the tagged polynucleotide fragments in the presence primers corresponding to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. Additionally, any combination of these genes can be amplified. For example, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, or all 54 of these genes can be amplified.

[00230] The methods described herein can comprise generating a plurality of sequence reads from a plurality of polynucleotide molecules. The plurality of polynucleotide molecules can cover genomic loci of a target genome. For example, the genomic loci can correspond to a plurality of genes as listed above. Further, the genomic loci can be any combination of these genes. Any given genomic locus can comprise at least two nucleic acid bases. Any given genomic locus can also comprise a plurality of nucleic acid bases, for example, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, or more nucleic acid bases.

[00231] The method can comprise grouping with a computer processor the plurality of sequence reads into families. Each of the family can comprises sequence reads from one of the template polynucleotides. Each family can comprise sequence reads from only one of the template polynucleotides. For each of the family, the sequence reads can be merged to generate a consensus sequence. The grouping can comprise classifying the plurality of sequence reads into families by identifying (i) distinct molecular barcodes coupled to the plurality of polynucleotide molecules and (ii) similarities between the plurality of sequence

reads, wherein each family includes a plurality of nucleic acid sequences that are associated with a distinct combination of molecular barcodes and similar or identical sequence reads.

[00232] Once merged, a consensus sequence can be called at a given genomic locus among the genomic loci. At any given genomic loci, any of the following can be determined: i) genetic variants among the calls; ii) frequency of a genetic alteration among the calls; iii) total number of calls; and iv) total number of alterations among the calls. The calling can comprise calling at least one nucleic acid base at the given genomic locus. The calling can also comprise calling a plurality of nucleic acid bases at the given genomic locus. In some cases, the calling can comprise phylogenetic analysis, voting (*e.g.*, biased voting), weighing, assigning a probability to each read at the locus in a family, or calling the base with the highest probability. The consensus sequence can be generated by evaluating a quantitative measure or a statistical significance level for each of the sequence reads. If a quantitative measure is performed, the method can comprise use of a binomial distribution, exponential distribution, beta distribution, or empirical distribution. However, frequency of the base at the particular location can also be used for calling, for example, if 51% or more of the reads is a "A" at the location, then the base may be called an "A" at that particular location. The method can further comprise mapping a consensus sequence to a target genome.

[00233] The method can further comprising performing consensus calling at an additional genomic locus among the genomic loci. The method can comprise determining a variation in copy number at one of the given genomic locus and additional genomic locus based on counts at the given genomic locus and additional genomic locus.

[00234] The methods described herein can comprise providing template polynucleotide molecules and a library of adaptor polynucleotide molecules in a reaction vessel. The adaptor polynucleotide molecules can have from 2 to 1,000 different barcode sequences and in some cases are not sequencing adaptors. Other variations of adaptor polynucleotide molecules are described throughout, which can also be used in the methods.

[00235] The polynucleotide molecules of the adaptors can have the same sample tag. The adaptor polynucleotide molecules can be coupled to both ends of the template polynucleotide molecules. The method can comprise coupling the adaptor polynucleotide molecules to the template polynucleotide molecules at an efficiency of at least 30%, *e.g.*, of at least 50% (*e.g.*, 60%, 70%, 80%, 90%, 95%, or 99%), thereby tagging each template polynucleotide with a tagging combination that is among 4 to 1,000,000 different tagging combinations, to produce tagged polynucleotide molecules. In some cases, the reaction can occur in a single reaction

vessel. Coupling efficiency can also be at least 30%, 35%, 40%, 45%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99%. Tagging can be non-unique tagging.

[00236] The tagged polynucleotide molecules can then be subject to an amplification reaction under conditions that will yield amplified polynucleotide molecules as amplification products of the tagged polynucleotide molecules. The template polynucleotide molecules can be double-stranded. Further, the template polynucleotide molecules can be blunt ended. In some cases, the amplification reaction comprises non-specifically amplifying the tagged polynucleotide molecules. The amplification reaction can also comprises using a priming site to amplify each of the tagged polynucleotide molecules. The priming site can be a primer, *e.g.*, a universal primer. The priming site can also be a nick.

[00237] The method can also comprise sequencing the amplified polynucleotide molecules. The sequencing can comprise (i) subjecting the amplified polynucleotide molecules to an additional amplification reaction under conditions that yield additional amplified polynucleotide molecules as amplification products of the amplified polynucleotide molecules, and/or (ii) sequencing the additional amplified polynucleotide molecules. The additional amplification can be performed in the presence of primers comprising flow cells sequences, which will produce polynucleotide molecules that are capable of binding to a flow cell. The additional amplification can also be performed in the presence of primers comprising sequences for hairpin shaped adaptors. The hairpin shaped adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times. The method can further comprise identifying genetic variants upon sequencing the amplified polynucleotide molecules.

[00238] The method can further comprising separating polynucleotide molecules comprising one or more given sequences from the amplified polynucleotide molecules, to produce enriched polynucleotide molecules. The method can also comprise amplifying the enriched polynucleotide molecules with primers comprising the flow cell sequences. This amplification with primers comprising flow cell sequences will produce polynucleotide molecules that are capable of binding to a flow cell. The amplification can also be performed in the presence of primers comprising sequences for hairpin shaped adaptors. The hairpin

shaped adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times.

[00239] Flow cell sequences or hairpin shaped adaptors can be added by non-amplification methods such as through ligation of such sequences. Other techniques such as hybridization methods can be used, *e.g.*, nucleotide overhangs.

[00240] The method can be performed without aliquoting the tagged polynucleotide molecules. For example, once the tagged polynucleotide molecule is made, the amplification and sequencing can occur in the same tube without any further preparation.

[00241] The methods described herein can be useful in detecting single nucleotide variations (SNV), copy number variations (CNV), insertions, deletions, and/or rearrangements. In some cases, the SNVs, CNVs, insertions, deletions, and/or rearrangements, can be associated with disease, for example, cancer.

[00242] **N. Monitoring a Patient's Status**

[00243] Methods disclosed herein can also be used to monitor a patient's disease status. The disease of a subject can be monitored over time to determine a progression of the disease (e.g., regression). Markers indicative of the disease can be monitored in a biological sample of the subject, such as a cell-free DNA sample.

[00244] For example, monitoring a subject's cancer status can comprise (a) determining an amount of one or more SNVs or copy numbers of a plurality of genes (e.g., in an exon), (b) repeating such determination at different points in time, and (c) determining if there is a difference in the number of SNVs, level of SNVs, number or level of genomic rearrangements, or copy numbers between (a) and (b). The genes can be selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. The genes can be selected from any 5, 10, 15, 20, 30, 40, 50, or all of the genes in this group.

[00245] O. Sensitivity and Specificity

[00246] Methods disclosed herein can be used to detect cancer polynucleotides in a sample, and cancer in a subject, with high measures of agreement, *e.g.*, high sensitivity and/or specificity. For example, such methods can detect cancer polynucleotides (e.g., rare DNA) in a sample at a concentration that is less than 5%, 1%, 0.5%, 0.1%, 0.05%, or 0.01%, at a specificity of at least 99%, 99.9%, 99.99%, 99.999%, 99.9999%, or 99.99999%. Such polynucleotides may be indicative of cancer or other disease. Further, such methods can detect cancer polynucleotides in a sample with a positive predictive value of at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 99.9%, 99.99%, 99.999%, or 99.9999%.

[00247] Subjects identified as positive in a test that are in reality positive are referred as true positives (TP). Subjects identified as positive in a test that are in reality negative are referred as false positives (FP). Subjects identified as negative in a test that are in reality negative are referred as true negatives (TN). Subjects identified as negative in a test that are in reality positive are referred as false negatives (FN). Sensitivity is the percentage of actual positives identified in a test as positive. This includes, for example, instances in which one should have found a cancer genetic variant and did. (Sensitivity =TP/(TP+FN).) Specificity is the percentage of actual negatives identified in a test as negative. This includes, for example, instances in which one should have found no cancer genetic variant and did not. Specificity can be calculated using the following equation: Specificity = TN/(TN+FP). Positive predictive value (PPV) can be measured by the percentage of subjects who test positive that are true positives. PPV can be calculated using the following equation: PPV= TP/(TP+FP). Positive predictive value can be increased by increasing sensitivity (*e.g.*, chance of an actual positive being detected) and/or specificity (*e.g.*, chance of not mistaking an actual negative for a positive).

[00248] Low conversion rates of polynucleotides into adaptor-tagged polynucleotides can compromise sensitivity as it decreases the chance of converting, and therefore detecting, rare polynucleotide targets. Noise in a test can compromise specificity as it increases the number of false positives detected in a test. Both low conversion rate and noise compromise positive predictive value as they decrease the percentage of true positives and increase the percentage of false positives.

[00249] The methods disclosed herein can achieve high levels of agreement, *e.g.*, sensitivity and specificity, leading to high positive predictive values. Methods of increasing sensitivity include high efficiency conversion of polynucleotides into adaptor-tagged

polynucleotides in a sample. Methods of increasing specificity include reducing sequencing errors, for example, by molecular tracking.

[00250] Methods of the present disclosure can be used to detect genetic variation in non-uniquely tagged initial starting genetic material (e.g., rare DNA) at a concentration that is less than 5%, 1%, 0.5%, 0.1%, 0.05%, or 0.01%, at a specificity of at least 99%, 99.9%, 99.99%, 99.999%, 99.9999%, or 99.99999%. In some aspects, the methods can further comprise converting polynucleotides in the initial starting material at an efficiency of at least at least 10%, at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 80%, or at least 90%. Sequence reads of tagged polynucleotides can be subsequently tracked to generate consensus sequences for polynucleotides with an error rate of no more than 2%, 1%, 0.1%, or 0.01%.

[00251] **2. Pooling Methods**

[00252] Disclosed herein are methods of detecting copy number variation and/or sequence variants at one or more genetic loci in a test sample. One embodiment is shown in **FIG. 8**. Typically, detecting copy number variation involves determining a quantitative measure (*e.g.*, an absolute or relative number) of polynucleotides mapping to a genetic locus of interest in a genome of a test sample, and comparing that number to a quantitative measure of polynucleotides mapping to that locus in a control sample. In certain methods, the quantitative measure is determined by comparing the number of molecules in the test sample that map to a locus of interest with a number of molecules in the test sample mapping to a reference sequence, *e.g.*, a sequence expected to be present at wild type ploidy number. In some examples, the reference sequence is HG19, build 37, or build 38. The comparison could involve, for example, determining a ratio. Then, this measure is compared with a similar measure determined in a control sample. So, for example, if a test sample has a ratio of 1.5:1 for locus of interest versus reference locus, and a control sample has a ratio of 1:1 for the same loci, one may conclude that the test sample exhibits polyploidy at the locus of interest.

[00253] When the test sample and the control sample are analyzed separately, the work flow can introduce distortions between final numbers in the control and test samples.

[00254] In one method disclosed herein (*e.g.*, flow chart 800), polynucleotides are provided from a test and a control sample (802). Polynucleotides in a test sample and those in a control sample are tagged with tags that identify the polynucleotides as originating from

the test or control sample (a source tag). (804.) The tag can be, for example, a polynucleotide sequence or barcode that unambiguously identifies the source.

[00255] The polynucleotides in each of the control and test samples also can be tagged with identifier tags that will be carried by all amplification progeny of a polynucleotide. Information from start and end sequences of a polynucleotide and identifier tags can identify sequence reads from polynucleotides amplified from an original parent molecule. Each molecule can be uniquely tagged compared with other molecules in the sample. Alternatively, each molecule need not be uniquely tagged compared with other molecules in the sample. That is, the number of different identifier sequences can be fewer than that the number of molecules in sample. By combining identifier information with start/stop sequence information, the probability of confusing two molecules having the same start/stop sequence is significantly diminished.

[00256] Number of different identifiers used to tag a nucleic acid (*e.g.*, cfDNA) can dependent on the number of different haploid genome equivalents. Different identifiers can be used to tag at least 2, least 10, least 100, least 200, least 300, least 400, least 500, least 600, least 700, least 800, least 900, least 1,000, least 2,000, least 3,000, least 4,000, least 5,000, least 6,000, least 7,000, least 8,000, least 9,000, least 10,000 or more different haploid genome equivalents. Accordingly, the number of different identifiers used to tag a nucleic acid sample, *e.g.*, cell-free DNA from 500 to 10,000 different haploid genome equivalents and be between any of 1, 2, 3, 4 and 5 and no more than 100, 90, 80, 70, 60, 50, 40 or 30. For example, the number of different identifier used to tag a nucleic acid sample from 500 to 10,000 different haploid genome equivalents can be 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100 or less.

[00257] Polynucleotides can be tagged by ligation of adaptors comprising the tags or identifiers before amplification. Ligation can be performed using an enzyme, *e.g.*, a ligase. For example, tagging can be performed using a DNA ligase. The DNA ligase can be a T4 DNA ligase, *E. coli* DNA ligase, and/or mammalian ligase. The mammalian ligase can be DNA ligase I, DNA ligase III, or DNA ligase IV. The ligase may also be a thermostable ligase. Tags can be ligated to a blunt-end of a polynucleotide (blunt-end ligation). Alternatively, tags can be ligated to a sticky end of a polynucleotide (sticky-end ligation).

The polynucleotides can be tagged by blunt end ligation using adaptors (*e.g.*, adaptors having forked ends). High efficiency of ligation can be achieved using high excess of adaptors (*e.g.*, more than 1.5X, more than 2X, more than 3X, more than 4X, more than 5X, more than 6X, more than 7X, more than 8X, more than 9X, more than 10X, more than 11X, more than 12X, more than 13X, more than 14X, more than 15X, more than 20X, more than 25X, more than 30X, more than 35X, more than 40X, more than 45X, more than 50X, more than 55X, more than 60X, more than 65X, more than 70X, more than 75X, more than 80X, more than 85X, more than 90X, more than 95X, or more than 100).

[00258] Once tagged with tags that identify source of the polynucleotides, polynucleotides from different sources (*e.g.*, different samples) can be pooled. After pooling, polynucleotides from different sources (*e.g.*, different samples) can be distinguished by any measurement using the tags, including any process of quantitative measurement. For example, as shown in (806) (**FIG. 8**), polynucleotides from the control sample and the test sample can be pooled. The pooled molecules can be subject to the sequencing (808) and bioinformatic work flow. Both will be subject to the same variations in the process and, therefore, any differential bias is reduced. Because molecules originating from control and test samples are differently tagged, they can be distinguished in any process of quantitative measurement.

[00259] The relative amount of control and test sample pooled can be varied. The amount of control sample can be same as the amount of test sample. The amount of control sample can also be larger than the amount of test sample. Alternatively, the amount of control sample can be smaller than the amount of test sample. The smaller the relative amount of one sample to the total, the fewer identifying tags needed in the original tagging process. A number can be selected to reduce to acceptable levels the probability that two parent molecules having the same start/end sequences will bear the same identifying tag. This probability can be less than 10%, less than 1%, less than 0.1% or less than 0.01%. The probability can be less than 25%, 24%, 23%, 22%, 21%, 20%, 19%, 18%, 17%, 16%, 15%, 14%, 13%, 12%, 11%, 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, or 1%.

[00260] Methods disclosed herein can also comprise grouping sequence reads. For example, bioinformatic workflow can include grouping sequence reads produced from progeny of a single parent molecule, as shown in (810) (**FIG. 8**). This can involve any of the redundancy reduction methods described herein. Molecules sourced from test and control samples can be differentiated based on source tags they carry (812). Molecules mapping to a target locus are quantified for both test-sourced and control-sourced molecules (812). This

can include the normalization methods discussed herein, *e.g.*, in which numbers at a target locus are normalized against numbers at a reference locus.

[00261] Normalized (or raw) quantities at a target locus from test and control samples are compared to determine presence of copy number variation (814).

[00262] **3. Computer Control Systems**

[00263] The present disclosure provides computer control systems that are programmed to implement methods of the disclosure. **FIG. 6** shows a computer system 1501 that is programmed or otherwise configured to implement the methods of the present disclosure. The computer system 1501 can regulate various aspects sample preparation, sequencing and/or analysis. In some examples, the computer system 1501 is configured to perform sample preparation and sample analysis, including nucleic acid sequencing. The computer system 1501 can be an electronic device of a user or a computer system that is remotely located with respect to the electronic device. The electronic device can be a mobile electronic device.

[00264] The computer system 1501 includes a central processing unit (CPU, also "processor" and "computer processor" herein) 1505, which can be a single core or multi core processor, or a plurality of processors for parallel processing. The computer system 1501 also includes memory or memory location 1510 (*e.g.*, random-access memory, read-only memory, flash memory), electronic storage unit 1515 (*e.g.*, hard disk), communication interface 1520 (*e.g.*, network adapter) for communicating with one or more other systems, and peripheral devices 1525, such as cache, other memory, data storage and/or electronic display adapters. The memory 1510, storage unit 1515, interface 1520 and peripheral devices 1525 are in communication with the CPU 1505 through a communication bus (solid lines), such as a motherboard. The storage unit 1515 can be a data storage unit (or data repository) for storing data. The computer system 1501 can be operatively coupled to a computer network ("network") 1530 with the aid of the communication interface 1520. The network 1530 can be the Internet, an internet and/or extranet, or an intranet and/or extranet that is in communication with the Internet. The network 1530 in some cases is a telecommunication and/or data network. The network 1530 can include one or more computer servers, which can enable distributed computing, such as cloud computing. The network 1530, in some cases with the aid of the computer system 1501, can implement a peer-to-peer network, which may enable devices coupled to the computer system 1501 to behave as a client or a server.

[00265] The CPU 1505 can execute a sequence of machine-readable instructions, which can be embodied in a program or software. The instructions may be stored in a memory location, such as the memory 1510. The instructions can be directed to the CPU 1505, which can subsequently program or otherwise configure the CPU 1505 to implement methods of the present disclosure. Examples of operations performed by the CPU 1505 can include fetch, decode, execute, and writeback.

[00266] The CPU 1505 can be part of a circuit, such as an integrated circuit. One or more other components of the system 1501 can be included in the circuit. In some cases, the circuit is an application specific integrated circuit (ASIC).

[00267] The storage unit 1515 can store files, such as drivers, libraries and saved programs. The storage unit 1515 can store user data, *e.g.*, user preferences and user programs. The computer system 1501 in some cases can include one or more additional data storage units that are external to the computer system 1501, such as located on a remote server that is in communication with the computer system 1501 through an intranet or the Internet.

[00268] The computer system 1501 can communicate with one or more remote computer systems through the network 1530. For instance, the computer system 1501 can communicate with a remote computer system of a user (*e.g.*, an operator). Examples of remote computer systems include personal computers (*e.g.*, portable PC), slate or tablet PC's (*e.g.*, Apple® iPad, Samsung® Galaxy Tab), telephones, Smart phones (*e.g.*, Apple® iPhone, Android-enabled device, Blackberry®), or personal digital assistants. The user can access the computer system 1501 via the network 1530.

[00269] Methods as described herein can be implemented by way of machine (*e.g.*, computer processor) executable code stored on an electronic storage location of the computer system 1501, such as, for example, on the memory 1510 or electronic storage unit 1515. The machine executable or machine readable code can be provided in the form of software. During use, the code can be executed by the processor 1505. In some cases, the code can be retrieved from the storage unit 1515 and stored on the memory 1510 for ready access by the processor 1505. In some situations, the electronic storage unit 1515 can be precluded, and machine-executable instructions are stored on memory 1510.

[00270] The code can be pre-compiled and configured for use with a machine have a processer adapted to execute the code, or can be compiled during runtime. The code can be

supplied in a programming language that can be selected to enable the code to execute in a pre-compiled or as-compiled fashion.

[00271] Aspects of the systems and methods provided herein, such as the computer system 1501, can be embodied in programming. Various aspects of the technology may be thought of as "products" or "articles of manufacture" typically in the form of machine (or processor) executable code and/or associated data that is carried on or embodied in a type of machine readable medium. Machine-executable code can be stored on an electronic storage unit, such memory (*e.g.*, read-only memory, random-access memory, flash memory) or a hard disk. "Storage" type media can include any or all of the tangible memory of the computers, processors or the like, or associated modules thereof, such as various semiconductor memories, tape drives, disk drives and the like, which may provide non-transitory storage at any time for the software programming. All or portions of the software may at times be communicated through the Internet or various other telecommunication networks. Such communications, for example, may enable loading of the software from one computer or processor into another, for example, from a management server or host computer into the computer platform of an application server. Thus, another type of media that may bear the software elements includes optical, electrical and electromagnetic waves, such as used across physical interfaces between local devices, through wired and optical landline networks and over various air-links. The physical elements that carry such waves, such as wired or wireless links, optical links or the like, also may be considered as media bearing the software. As used herein, unless restricted to non-transitory, tangible "storage" media, terms such as computer or machine "readable medium" refer to any medium that participates in providing instructions to a processor for execution.

[00272] Hence, a machine readable medium, such as computer-executable code, may take many forms, including but not limited to, a tangible storage medium, a carrier wave medium or physical transmission medium. Non-volatile storage media include, for example, optical or magnetic disks, such as any of the storage devices in any computer(s) or the like, such as may be used to implement the databases, etc. shown in the drawings. Volatile storage media include dynamic memory, such as main memory of such a computer platform. Tangible transmission media include coaxial cables; copper wire and fiber optics, including the wires that comprise a bus within a computer system. Carrier-wave transmission media may take the form of electric or electromagnetic signals, or acoustic or light waves such as those generated during radio frequency (RF) and infrared (IR) data communications. Common

forms of computer-readable media therefore include for example: a floppy disk, a flexible disk, hard disk, magnetic tape, any other magnetic medium, a CD-ROM, DVD or DVD-ROM, any other optical medium, punch cards paper tape, any other physical storage medium with patterns of holes, a RAM, a ROM, a PROM and EPROM, a FLASH-EPROM, any other memory chip or cartridge, a carrier wave transporting data or instructions, cables or links transporting such a carrier wave, or any other medium from which a computer may read programming code and/or data. Many of these forms of computer readable media may be involved in carrying one or more sequences of one or more instructions to a processor for execution.

[00273] The computer system 1501 can include or be in communication with an electronic display 1535 that comprises a user interface (UI) 1540. The UI can allow a user to set various conditions for the methods described herein, for example, PCR or sequencing conditions. Examples of UI's include, without limitation, a graphical user interface (GUI) and web-based user interface.

[00274] Methods and systems of the present disclosure can be implemented by way of one or more algorithms. An algorithm can be implemented by way of software upon execution by the central processing unit 1505. The algorithm can, for example, process the reads to generate a consequence sequence.

[00275] **FIG. 7** schematically illustrates another system for analyzing a sample comprising nucleic acids from a subject. The system includes a sequencer, bioinformatic software and internet connection for report analysis by, for example, a hand held device or a desktop computer

[00276] Disclosed herein is a system for analyzing a target nucleic acid molecule of a subject, comprising: a communication interface that receives nucleic acid sequence reads for a plurality of polynucleotide molecules that cover genomic loci of a target genome; computer memory that stores the nucleic acid sequence reads for the plurality of polynucleotide molecules received by the communication interface; and a computer processor operatively coupled to the communication interface and the memory and programmed to (i) group the plurality of sequence reads into families, wherein each family comprises sequence reads from one of the template polynucleotides, (ii) for each of the families, merge sequence reads to generate a consensus sequence, (iii) call the consensus sequence at a given genomic locus among the genomic loci, and (iv) detect at the given genomic locus any of genetic variants among the calls, frequency of a genetic alteration among the calls, total number of calls; and

total number of alterations among the calls, wherein the genomic loci correspond to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. The different variations of each component of the system are described throughout the disclosure within the methods and compositions. These individual components and variations thereof, are also applicable in this system.

[00277] **4. Kits**

[00278] Kits comprising the compositions as described herein. The kits can be useful in performing the methods as described herein. Disclosed herein is a kit comprising a plurality of oligonucleotide probes that selectively hybridize to least 5, 6, 7, 8, 9, 10, 20, 30, 40 or all genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. The number genes to which the oligonucleotide probes can selectively hybridize can vary. For example, the number of genes can comprise 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, or 54. The kit can include a container that includes the plurality of oligonucleotide probes and instructions for performing any of the methods described herein.

[00279] The oligonucleotide probes can selectively hybridize to exon regions of the genes, *e.g.*, of the at least 5 genes. In some cases, the oligonucleotide probes can selectively hybridize to at least 30 exons of the genes, *e.g.*, of the at least 5 genes. In some cases, the multiple probes can selectively hybridize to each of the at least 30 exons. The probes that

hybridize to each exon can have sequences that overlap with at least 1 other probe. In some embodiments, the oligoprobes can selectively hybridize to non-coding regions of genes disclosed herein, for example, intronic regions of the genes. The oligoprobes can also selectively hybridize to regions of genes comprising both exonic and intronic regions of the genes disclosed herein.

[00280] Any number of exons can be targeted by the oligonucleotide probes. For example, at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, 145, 150, 155, 160, 165, 170, 175, 180, 185, 190, 195, 200, 205, 210, 215, 220, 225, 230, 235, 240, 245, 250, 255, 260, 265, 270, 275, 280, 285, 290, , 295, 300, 400, 500, 600, 700, 800, 900, 1,000, or more, exons can be targeted.

[00281] The kit can comprise at least 4, 5, 6, 7, or 8 different library adaptors having distinct molecular barcodes and identical sample barcodes. The library adaptors may not be sequencing adaptors. For example, the library adaptors do not include flow cell sequences or sequences that permit the formation of hairpin loops for sequencing. The different variations and combinations of molecular barcodes and sample barcodes are described throughout, and are applicable to the kit. Further, in some cases, the adaptors are not sequencing adaptors. Additionally, the adaptors provided with the kit can also comprise sequencing adaptors. A sequencing adaptor can comprise a sequence hybridizing to one or more sequencing primers. A sequencing adaptor can further comprise a sequence hybridizing to a solid support, *e.g.*, a flow cell sequence. For example, a sequencing adaptor can be a flow cell adaptor. The sequencing adaptors can be attached to one or both ends of a polynucleotide fragment. In some cases, the kit can comprise at least 8 different library adaptors having distinct molecular barcodes and identical sample barcodes. The library adaptors may not be sequencing adaptors. The kit can further include a sequencing adaptor having a first sequence that selectively hybridizes to the library adaptors and a second sequence that selectively hybridizes to a flow cell sequence. In another example, a sequencing adaptor can be hairpin shaped. For example, the hairpin shaped adaptor can comprise a complementary double stranded portion and a loop portion, where the double stranded portion can be attached (*e.g.*, ligated) to a double-stranded polynucleotide. Hairpin shaped sequencing adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times. A sequencing adaptor can be up to 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41,

42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, or more bases from end to end. The sequencing adaptor can comprise 20-30, 20-40, 30-50, 30-60, 40-60, 40-70, 50-60, 50-70, bases from end to end. In a particular example, the sequencing adaptor can comprise 20-30 bases from end to end. In another example, the sequencing adaptor can comprise 50-60 bases from end to end. A sequencing adaptor can comprise one or more barcodes. For example, a sequencing adaptor can comprise a sample barcode. The sample barcode can comprise a pre-determined sequence. The sample barcodes can be used to identify the source of the polynucleotides. The sample barcode can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or more (or any length as described throughout) nucleic acid bases, *e.g.*, at least 8 bases. The barcode can be contiguous or non-contiguous sequences, as described above.

[00282] The library adaptors can be blunt ended and Y-shaped and can be less than or equal to 40 nucleic acid bases in length. Other variations of the can be found throughout and are applicable to the kit.

EXAMPLES

[00283] **Example 1. Methods for copy number variation detection.**

[00284] Blood collection

[00285] 10-30 mL Blood samples are collected at room temperature. The samples are centrifuged to remove cells. Plasma is collected after centrifugation.

[00286] cfDNA extraction

[00287] The sample is subjected to proteinase K digestion. DNA is precipitated with isopropanol. DNA is captured on a DNA purification column (*e.g.*, a QIAamp DNA Blood Mini Kit) and eluted in 100 µl solution. DNAs below 500 bp are selected with Ampure SPRI magnetic bead capture (PEG/salt). The resulting production is suspended in 30 µl H_2O. Size distribution is checked (major peak = 166 nucleotides; minor peak = 330 nucleotides) and quantified. 5 ng of extracted DNA contain approximately 1700 haploid genome equivalents ("HGE"). The general correlation between the amount of DNA and HGE is as follow: 3 pg DNA = 1 HGE; 3 ng DNA = 1K HGE; 3 µg DNA = 1M HGE; 10 pg DNA = 3 HE; 10 ng DNA = 3K HGE; 10 µg DNA = 3M HGE.

[00288] "Single Molecule" library prep

[00289] High-efficiency DNA tagging (>80%) is performed by blunt-end repair and ligation with 8 different octomers (*i.e.*, 64 combinations) with overloaded hairpin adaptors. 2.5 ng DNA (*i.e.* approximately 800 HGE) is used as the starting material. Each hairpin adaptor comprises a random sequence on its non-complementary portion. Both ends of each DNA fragment are attached with hairpin adaptors. Each tagged fragment can be identified by the random sequence on the hairpin adaptors and a 10 p endogenous sequence on the fragment.

[00290] Tagged DNA is amplified by 10 cycles of PCR to produce about 1-7 μg DNAs that contain approximately 500 copies of each of the 800 HGE in the starting material.

[00291] Buffer optimization, polymerase optimization and cycle reduction may be performed to optimize the PCR reactions. Amplification bias, *e.g.*, non-specific bias, GC bias, and/or size bias are also reduced by optimization. Noise(s) (*e.g.*, polymerase-introduced errors) are reduced by using high-fidelity polymerases.

[00292] The Library may be prepared using Verniata or Sequenom methods.

[00293] Sequences may be enriched as follow: DNAs with regions of interest (ROI) are captured using biotin-labeled bead with probe to ROIs. The ROIs are amplified with 12 cycles of PCR to generate a 2000 times amplification. The resulting DNA is then denatured and diluted to 8 pM and loaded into an Illumina sequencer.

[00294] Massively parallel sequencing

[00295] 0.1 to 1% of the sample (approximately 100pg) are used for sequencing.

[00296] Digital bioinformatics

[00297] Sequence reads are grouped into families, with about 10 sequence reads in each family. Families are collapsed into consensus sequences by voting (*e.g.*, biased voting) each position in a family. A base is called for consensus sequence if 8 or 9 members agree. A base is not called for consensus sequence if no more than 60% of the members agree.

[00298] The resulting consensus sequences are mapped to a reference genome. Each base in a consensus sequence is covered by about 3000 different families. A quality score for each sequence is calculated and sequences are filtered based on the their quality scores.

[00299] Sequence variation is detected by counting distribution of bases at each locus. If 98% of the reads have the same base (homozygous) and 2% have a different base, the locus is likely to have a sequence variant, presumably from cancer DNA.

[00300] CNV is detected by counting the total number of sequences (bases) mapping to a locus and comparing with a control locus. To increase CNV detection, CNV analysis is performed specific regions, including regions on ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, or NTRK1 genes.

[00301] Example 2. Method for Correcting Base Calling by Determining the Total Number Unseen Molecules in a Sample

[00302] After fragments are amplified and the sequences of amplified fragments are read and aligned, the fragments are subjected to base calling. Variations in the number of amplified fragments and unseen amplified fragments can introduce errors in base calling. These variations are corrected by calculating the number of unseen amplified fragments.

[00303] When base calling for locus A (an arbitrary locus), it is first assumed that there are N amplified fragments. The sequence readouts can come from two types of fragments: double-strand fragments and single-strand fragments. The following is a theoretical example of calculating the total number of unseen molecules in a sample.

[00304] N is the total number of molecules in the sample.
Assuming 1000 is the number of duplexes detected.
Assuming 500 is the number of single-stranded molecule detected.
P is the probability of seeing a strand.
Q is the probability of not detecting a strand.

[00305] Since Q = 1 -P.
1000 = NP(2).
500 = N2PQ.
1000 / P(2) = N.
500÷2 PQ = N.
1000 / P(2) = 500÷2PQ.
1000 * 2 PQ = 500 P(2).
2000 PQ = 500 P(2).
2000 Q = 500 P.
2000 (1-P) = 500P
2000-2000 P = 500P.
2000 = 500P +2000 P.
2000 = 2500 P.

2000÷2500 = P.
0.8 = P.
1000/ P(2) = N.
1000÷0.64 = N.
1562 = N.
Number of unseen fragments = 62.

[00306] Example 3. Identification of genetic variants in cancer-associated somatic variants in a patient.

[00307] An assay is used to analyze a panel of genes to identify genetic variants in cancer-associated somatic variants with high sensitivity.

[00308] Cell-free DNA is extracted from plasma of a patient and amplified by PCR. Genetic variants are analyzed by massively parallel sequencing of the amplified target genes. For one set of genes, all exons are sequenced as such sequencing coverage had shown to have clinically utility (Table 1). For another set of genes, sequencing coverage included those exons with a previously reported somatic mutation (Table 2). The minimum detectable mutant allele (limit of detection) is dependent on the patient's sample cell-free DNA concentration, which varied from less than 10 to over 1,000 genomic equivalents per mL of peripheral blood. Amplification may not be detected in samples with lower amounts of cell-free DNA and/or low-level gene copy amplification. Certain sample or variant characteristics resulted in reduced analytic sensitivity, such as low sample quality or improper collection.

[00309] The percentage of genetic variants found in cell-free DNA circulating in blood is related to the unique tumor biology of this patient. Factors that affected the amount/percentages of detected genetic variants in circulating cell-free DNA in blood include tumor growth, turn-over, size, heterogeneity, vascularization, disease progression or treatment. Table 3 annotates the percentage, or allele frequency, of altered circulating cell-free DNA (% cfDNA) detected in this patient. Some of the detected genetic variants are listed in descending order by % cfDNA.

[00310] Genetic variants are detected in the circulating cell-free DNA isolated from this patient's blood specimen. These genetic variants are cancer-associated somatic variants, some of which have been associated with either increased or reduced clinical response to specific treatment. "Minor Alterations" are defined as those alterations detected at less than 10% the allele frequency of "Major Alterations". The detected allele frequencies of these alterations (Table 3) and associated treatments for this patient are annotated.

[00311] All genes listed in Tables 1 and 2 are analyzed as part of the Guardant360™ test. Amplification is not detected for *ERBB2*, *EGFR*, or *MET* in the circulating cell-free DNA isolated from this patient's blood specimen.

[00312] Patient test results comprising the genetic variants are listed in Table 4.

Table 1. Genes in which all exons are sequenced

GENES IN WHICH ALL EXONS ARE SEQUENCED			
ALK	< 0.1%	APC	< 0.1%
AR	< 0.1%	BRAF	< 0.1%
CDKN2A	< 0.1%	EGFR	< 0.1%
ERBB2	< 0.1%	FBXW7	< 0.1%
KRAS	< 0.1%	MET	< 0.1%
MYC	< 0.1%	NOTCH1	< 0.1%
NRAS	< 0.1%	PIK3CA	< 0.1%
PTEN	< 0.1%	PROC	< 0.1%
RB1	< 0.1%	TP53	< 0.1%

LOD: Limit of Detection. The minimum detectable mutant allele frequency for this specimen in which 80% of somatic variants is detected.

Table 2. Genes in which exons with a previously reported somatic mutation are sequenced

GENES IN WHICH EXONS WITH A PREVIOUSLY REPORTED SOMATIC MUTATION ARE SEQUENCED			
ABL1	< 0.1%	AKT1	< 0.1%
ATM	< 0.1%	CDH1	< 0.1%
CSF1R	< 0.1%	CTNNB1	< 0.1%
ERBB4	< 0.1%	EZH2	< 0.1%
FGFR1	< 0.1%	FGFR2	< 0.1%
FGFR3	< 0.1%	FLT3	< 0.1%
GNA11	< 0.1%	GNAQ	< 0.1%
GNAS	< 0.1%	HNF1A	< 0.1%
HRAS	< 0.1%	IDH1	< 0.1%
IDH2	< 0.1%	JAK2	< 0.1%
JAK3	< 0.1%	KDR	< 0.1%
KIT	< 0.1%	MLH1	< 0.1%
MPL	< 0.1%	NPM1	< 0.1%
PDGFRA	< 0.1%	PTPN11	< 0.1%
RET	< 0.1%	SMAD4	< 0.1%
SMARCB1	< 0.1%	SMO	< 0.1%
SRC	< 0.1%	STK11	< 0.1%
TERT	< 0.1%	VHL	< 0.1%

LOD: Limit of Detection. The minimum detectable mutant allele frequency for this specimen in which 80% of somatic variants is detected.

Table 3. Allele frequency of altered circulating cell-free DNA detected in this patient

Gene	% cfDNA	0.01% — 0.1% — 1% — 10% — 100%
BRAF V600E	8.9%	8.9% / 91.1%
NRAS Q61K	6.2%	6.2% / 93.8%
JAK2 V617F	1.5%	1.5% / 98.6%

Legend:
% cfDNA : Allele frequency of genomic alteration observed in this patient's circulating cell-free DNA.
: Cell-free DNA with alterations.
: Cell-free DNA without alterations.

Table 4. Genomic alterations detected in selected genes

Detected: 51 Genomic Alterations

Gene	Chromosome	Position	Mutation (nt)	Mutation (AA)	Percentage	Cosmic ID	DBSNP ID
KRAS	12	25368462	C>T		100.0%		rs4362222
ALK	2	29416572	T>C	I1461V	100.0%		rs1670283
ALK	2	29444095	C>T		100.0%		rs1569156
ALK	2	29543663	T>C	Q500Q	100.0%		rs2293564
ALK	2	29940529	A>T	P234P	100.0%		rs2246745
APC	5	112176756	T>A	V1822D	100.0%		rs459552
CDKN2A	9	21968199	C>G		100.0%	COSM14251	rs11515
FGFR3	4	1807894	G>A	T651T	100.0%		rs7688609
NOTCH1	9	139410424	A>G		100.0%		rs3125006
PDGFRA	4	55141055	A>G	P567P	100.0%		rs1873778
HRAS	11	534242	A>G	H27H	100.0%	COSM249860	rs12628
EGFR	7	55214348	C>T	N158N	99.9%	COSM42978	rs2072454
TP53	17	7579472	G>C	P72R	99.8%		rs1042522
APC	5	112162854	T>C	Y486Y	55.0%		rs2229992
APC	5	112177171	G>A	P1960P	53.8%		rs465899
EGFR	7	55266417	T>C	T903T	53.6%		rs1140475
APC	5	112176325	G>A	G1678G	53.2%		rs42427
APC	5	112176559	T>G	S1756S	53.0%		rs866006
EGFR	7	55229255	G>A	R521K	53.0%		
MET	7	116397572	A>G	Q648Q	52.7%		
APC	5	112175770	G>A	T1493T	52.7%		rs41115
EGFR	7	55249063	G>A	Q787Q	52.6%		rs1050171
NOTCH1	9	139411714	T>C		52.4%		rs11145767
EGFR	7	55238874	T>A	T629T	52.0%		rs2227984
ERBB2	17	37879588	A>G	I655V	51.6%		rs1136201
NOTCH1	9	139397707	G>A	D1698D	51.3%	COSM33747	rs10521
ALK	2	30143499	G>C	L9L	51.0%		rs4358080
APC	5	112164561	G>A	A545A	51.0%		rs351771
FLT3	13	28610183	A>G		50.8%		rs2491231
NOTCH1	9	139418260	A>G	N104N	50.5%		rs4489420
ALK	2	29444076	G>T		50.4%		rs1534545
PIK3CA	3	178917005	A>G		50.3%		rs3729674
NOTCH1	9	139412197	G>A		50.2%		rs9411208
ALK	2	29455267	A>G	G845G	50.0%	COSM148825	rs2256740
KIT	4	55593464	A>C	M541L	49.9%	COSM28026	
NOTCH1	9	139391636	G>A	D2185D	48.9%		rs2229974
PDGFRA	4	55152040	C>T	V824V	48.9%	COSM22413	rs2228230
ALK	2	29416481	T>C	K1491R	48.9%	COSM1130802	rs1881420
ALK	2	29445458	G>T	G1125G	48.6%		rs3795850
NOTCH1	9	139410177	T>C		48.5%		rs3124603

RET	10	43613843	G>T	L769L	48.2%		rs1800861
EGFR	7	55214443	G>A		48.0%		rs7801956
ALK	2	29416366	G>C	D1529E	47.2%		rs1881421
EGFR	7	55238087	C>T		45.5%		rs10258429
RET	10	43615633	C>G	S904S	44.8%		rs1800863
BRAF	7	140453136	A>T	V600E	8.9%	COSM476	
NRAS	1	115256530	G>T	Q61K	6.2%	COSM580	rs121913254
JAK2	9	5073770	G>T	V617F	1.5%	COSM12600	rs77375493

[00313] Example 4. Determining patient-specific limits of detection for genes analyzed by Guardant360™ assays.

[00314] Using the method of Example 3, Genetic alterations in cell-free DNA of a patient are detected. The sequence reads of these genes include exon and/or intron sequences.

[00315] Limits of detection of the test are shown in Table 5. The limits of detection values are dependent on cell-free DNA concentration and sequencing coverage for each gene.

Table 5. Limits of Detection of selected genes in a patient using Guardant

Complete Exon and Partial Intron Coverage					
APC	0.1%	AR *	0.2%	ARID1A	
BRAF *	0.1%	BRCA1		BRCA2	
CCND1 *		CCND2 *		CCNE1 *	
CDK4 *		CDK6 *		CDKN2A	0.1%
CDKN2B		EGFR *	< 0.1%	ERBB2 *	0.1%
FGFR1 *	< 0.1%	FGFR2 *	0.1%	HRAS	0.1%
KIT *	0.1%	KRAS *	0.1%	MET *	0.1%
MYC *	0.1%	NF1		NRAS	0.1%
PDGFRA *	0.1%	PIK3CA*	0.1%	PTEN	0.1%
RAF1 *		TP53	0.1%		
Exons Covered with Reported Somatic Mutations					
AKT1	0.1%	ALK	< 0.1%	ARAF	
ATM	0.1%	CDH1	0.1%	CTNNB1	0.1%
ESR1		EZH2	0.1%	FBXW7	0.1%
FGFR3	0.1%	GATA3		GNA11	0.1%
GNAQ	0.1%	GNAS	0.1%	HNF1A	0.1%
IDH1	0.1%	IDH2	0.1%	JAK2	0.1%
JAK3	0.1%	MAP2K1		MAP2K2	
MLH1	0.1%	MPL	0.2%	NFE2L2	
NOTCH1	0.1%	NPM1	0.1%	PTPN11	0.1%
RET	0.1%	RHEB		RHOA	
RIT1		ROS1		SMAD4	0.1%
SMO	0.1%	SRC	< 0.1%	STK11	0.2%
TERT	0.1%	VHL	0.2%		
Fusions					
ALK	< 0.1%	RET	0.1%	ROS1	
NTRK1					

LOD: Limit of Detection. The minimum detectable mutant allele frequency for this specimen in which 80% of somatic variants is detected. * indicates CNV genes.

[00316] Example 5. Correcting Sequence Errors Comparing Watson and Crick Sequences

[00317] Double-stranded cell-free DNA is isolated from the plasma of a patient. The cell-free DNA fragments are tagged using 16 different bubble-containing adaptors, each of which comprises a distinctive barcode. The bubble-containing adaptors are attached to both ends of each cell-free DNA fragment by ligation. After ligation, each of the cell-free DNA fragment can be distinctly identified by the sequence of the distinct barcodes and two 20 bp endogenous sequences at each end of the cell-free DNA fragment.

[00318] The tagged cell-free DNA fragments are amplified by PCR. The amplified fragments are enriched using beads comprising oligonucleotide probes that specifically bind to a group of cancer-associated genes. Therefore, cell-free DNA fragments from the group of cancer-associated genes are selectively enriched.

[00319] Sequencing adaptors, each of which comprises a sequencing primer binding site, a sample barcode, and a cell-flow sequence, are attached to the enriched DNA molecules. The resulting molecules are amplified by PCR.

[00320] Both strands of the amplified fragments are sequenced. Because each bubble-containing adaptor comprises a non-complementary portion (*e.g.*, the bubble), the sequence of the one strand of the bubble-containing adaptor is different from the sequence of the other strand (complement). Therefore, the sequence reads of amplicons derived from the Watson strand of an original cell-free DNA can be distinguished from amplicons from the Crick strand of the original cell-free DNA by the attached bubble-containing adaptor sequences.

[00321] The sequence reads from a strand of an original cell-free DNA fragment are compared to the sequence reads from the other strand of the original cell-free DNA fragment. If a variant occurs in only the sequence reads from one strand, but not other strand, of the original cell-free DNA fragment, this variant will be identified as an error (*e.g.*, resulted from PCR and/or amplification), rather than a true genetic variant.

[00322] The sequence reads are grouped into families. Errors in the sequence reads are corrected. The consensus sequence of each family is generated by collapsing.

[00323] While preferred embodiments of the present invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. It is not intended that the invention be limited by the specific examples provided within the specification. While the invention has been described with reference to the aforementioned specification, the descriptions and illustrations of the

embodiments herein are not meant to be construed in a limiting sense. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the invention. Furthermore, it shall be understood that all aspects of the invention are not limited to the specific depictions, configurations or relative proportions set forth herein which depend upon a variety of conditions and variables. It should be understood that various alternatives to the embodiments of the invention described herein may be employed in practicing the invention. It is therefore contemplated that the invention shall also cover any such alternatives, modifications, variations or equivalents. It is intended that the following claims define the scope of the invention and that methods and structures within the scope of these claims and their equivalents be covered thereby.

CLAIMS

WHAT IS CLAIMED IS:

1. A method for determining a quantitative measure indicative of a number of individual double-stranded deoxyribonucleic acid (DNA) molecules in a sample, comprising:

(a) determining a quantitative measure of individual DNA molecules for which both strands are detected;

(b) determining a quantitative measure of individual DNA molecules for which only one of the DNA strands is detected;

(c) inferring from (a) and (b) above a quantitative measure of individual DNA molecules for which neither strand was detected; and

(d) using (a)-(c) to determine the quantitative measure indicative of a number of individual double-stranded DNA molecules in the sample.

2. The method of Claim 1, further comprising detecting copy number variation in said sample by determining a normalized quantitative measure determined in step (d) at each of one or more genetic loci and determining copy number variation based on the normalized measure.

3. The method of Claim 1, wherein said sample comprises double-stranded polynucleotide molecules sourced substantially from cell-free nucleic acids.

4. The method of Claim 1, wherein determining said quantitative measure of individual DNA molecules comprises tagging said DNA molecules with a set of duplex tags, wherein each duplex tag differently tags complementary strands of a double-stranded DNA molecule in said sample to provide tagged strands.

5. The method of Claim 4, further comprising sequencing at least some of said tagged strands to produce a set of sequence reads.

6. The method of Claim 5, further comprising sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in said set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among said sequence reads in said set of sequence reads.

7. The method of Claim 6, further comprising determining quantitative measures of (i) said paired reads and (ii) said unpaired reads that map to each of one or more genetic loci to determine a quantitative measure of total double-stranded DNA molecules in said sample that map to each of said one or more genetic loci based on said quantitative measure of paired reads and unpaired reads mapping to each locus.

8. A method for reducing distortion in a sequencing assay, comprising:
   (a) tagging control parent polynucleotides with a first tag set to produce tagged control parent polynucleotides;
   (b) tagging test parent polynucleotides with a second tag set to produce tagged test parent polynucleotides;
   (c) mixing tagged control parent polynucleotides with tagged test parent polynucleotides to form a pool;
   (d) determining quantities of tagged control parent polynucleotides and tagged test parent polynucleotides; and
   (e) using the quantities of tagged control parent polynucleotides to reduce distortion in the quantities of tagged test parent polynucleotides.

9. The method of Claim 8, wherein said first tag set comprises a plurality of tags, wherein each tag in said first tag set comprises a same control tag and an identifying tag, and wherein said first tag set comprises a plurality of different identifying tags.

10. The method of Claim 9, wherein said second tag set comprises a plurality of tags, wherein each tag in said second tag set comprises a same test tag and an identifying tag, wherein said test tag is distinguishable from said control tag, and wherein said second tag set comprises a plurality of different identifying tags.

11. The method of Claim 9, wherein (d) comprises amplifying tagged parent polynucleotides in said pool to form a pool of amplified, tagged polynucleotides, and sequencing amplified, tagged polynucleotides in said amplified pool to produce a plurality of sequence reads.

12. The method of Claim 11, further comprising grouping sequence reads into families, each family comprising sequence reads generated from a same parent polynucleotide, which grouping is optionally based on information from an identifying tag and from start/end

sequences of said parent polynucleotides, and, optionally, determining a consensus sequence for each of a plurality of parent polynucleotides from said plurality of sequence reads in a group.

13. The method of Claim 8, wherein (d) comprises determining copy number variation in said test parent polynucleotides at greater than or equal to one locus based on relative quantity of test parent polynucleotides and control parent polynucleotides mapping to said locus.

14. A set of library adaptors comprising a plurality of polynucleotide molecules with molecular barcodes, wherein said plurality of polynucleotide molecules are less than or equal to 80 nucleotide bases in length, wherein said molecular barcodes are at least 4 nucleotide bases in length, and wherein:

(a) said molecular barcodes are different from one another and have an edit distance of at least 1 between one another;

(b) said molecular barcodes are located at least one nucleotide base away from a terminal end of their respective polynucleotide molecules;

(c) optionally, at least one terminal base is identical in all of said polynucleotide molecules; and

(d) none of said polynucleotide molecules contains a complete sequencer motif.

15. The set of library adaptors of Claim 14, wherein said polynucleotide molecules are identical but for said molecular barcodes.

16. The set of library adaptors of Claim 14, wherein each of said plurality of polynucleotide molecules has a double stranded portion and at least one single-stranded portion.

17. The set of library adaptors of Claim 16, wherein said double-stranded portion has a molecular barcode among said molecular barcodes.

18. The set of library adaptors of Claim 17, wherein said given molecular barcode is a randomer.

19. The set of library adaptors of Claim 16, wherein each of said plurality of polynucleotide molecules further comprises a strand-identification barcode on said at least one single-stranded portion.

20. The set of library adaptors of Claim 19, wherein said strand-identification barcode includes at least 4 nucleotide bases.

21. The set of library adaptors of Claim 16, wherein said single-stranded portion has a partial sequencer motif.

22. The set of library adaptors of Claim 14, wherein said polynucleotide molecules have a sequence of terminal nucleotides that are the same.

23. The set of library adaptors of Claim 14, wherein each of said plurality of polynucleotide molecules is Y-shaped, bubble shaped or hairpin shaped.

24. The set of library adaptors of Claim 14, wherein none of said polynucleotide molecules contains a sample identification motif.

25. The set of library adaptors of Claim 14, wherein said molecular barcodes are at least 10 nucleotide bases in length.

26. The set of library adaptors of Claim 14, wherein each of said plurality of polynucleotide molecules is from 10 nucleotide bases to 60 nucleotide bases in length.

27. The set of library adaptors of Claim 14, where said at least one terminal base is identical in all of said polynucleotide molecules.

28. The set of library adaptors of Claim 14, wherein said molecular barcodes are located at least 10 nucleotide base away from a terminal end of their respective polynucleotide molecules.

29. The set of library adaptors of Claim 14, consisting essentially of said plurality of polynucleotide molecules.

30. A method, comprising:

(a) tagging a collection of polynucleotides with a plurality of polynucleotide molecules from a library of adaptors as in Claim 14 to create a collection of tagged polynucleotides; and

(b) amplifying said collection of tagged polynucleotides in the presence of sequencing adaptors, wherein said sequencing adaptors have primers with nucleotide sequences that are selectively hybridizable to complementary sequences in said plurality of polynucleotide molecules.

31. A method for detecting or quantifying rare deoxyribonucleic acid (DNA) in a heterogeneous population of original DNA fragments, wherein said rare DNA has a concentration that is less than 1%, the method comprising:

(a) tagging said original DNA fragments in a single reaction such that greater than 30% of said original DNA fragments are tagged at both ends with library adaptors that comprise molecular barcodes, thereby providing tagged DNA fragments;

(b) performing high-fidelity amplification on said tagged DNA fragments;

(c) optionally, selectively enriching a subset of said tagged DNA fragments;

(d) sequencing one or both strands of said tagged, amplified and optionally selectively enriched DNA fragments to obtain sequence reads comprising nucleotide sequences of said molecular barcodes and at least a portion of said original DNA fragments;

(e) from said sequence reads, determining consensus reads that are representative of single-strands of said original DNA fragments; and

(f) quantifying said consensus reads to detect or quantify said rare DNA at a specificity that is greater than 99.9%.

32. The method of Claim 31, wherein step (e) comprises comparing sequence reads having the same or similar molecular barcodes and the same or similar end of fragment sequences.

33. The method of Claim 32, wherein said comparing further comprises performing a phylogentic analysis on said sequence reads having the same or similar molecular barcodes.

34. The method of Claim 32, wherein said molecular barcodes include a barcode having an edit distance of up to 3.

35. The method of Claim 31, wherein said end of fragment sequence includes fragment sequences having an edit distance of up to 3.

36. The method of Claim 31, further comprising sorting sequence reads into paired reads and unpaired reads, and quantifying a number of paired reads and unpaired reads that map to each of one or more genetic loci.

37. The method of Claim 31, wherein said tagging occurs by having an excess amount of library adaptors as compared to original DNA fragments.

38. The method of Claim 31, further comprising binning said sequence reads according to said molecular barcodes and sequence information from at least one end of each of said original DNA fragments to create bins of single stranded reads.

39. The method of Claim 38, further comprising, in each bin, determining a sequence of a given original DNA fragment among said original DNA fragments by analyzing sequence reads.

40. The method of Claim 39, further comprising detecting or quantifying said rare DNA by comparing a number of times each base occurs at each position of a genome represented by said tagged, amplified, and optionally enriched DNA fragments.

41. The method of Claim 31, further comprising selectively enriching a subset of said tagged DNA fragments.

42. The method of Claim 41, further comprising, after enriching, amplifying the enriched tagged DNA fragments in the presence of sequencing adaptors comprising primers.

43. The method of Claim 31, wherein said DNA fragments are tagged with polynucleotide molecules from a library of adaptors as in Claim 1.

44. A method for processing and/or analyzing a nucleic acid sample of a subject, comprising:

(a) exposing polynucleotide fragments from said nucleic acid sample to a set of library adaptors to generate tagged polynucleotide fragments; and

(b) subjecting said tagged polynucleotide fragments to nucleic acid amplification reactions under conditions that yield amplified polynucleotide fragments as amplification products of said tagged polynucleotide fragments,

wherein said set of library adaptors comprises a plurality of polynucleotide molecules with molecular barcodes, wherein said plurality of polynucleotide molecules are less than or equal to 80 nucleotide bases in length, wherein said molecular barcodes are at least 4 nucleotide bases in length, and wherein

(1) said molecular barcodes are different from one another and have an edit distance of at least 1 between one another;

(2) said molecular barcodes are located at least one nucleotide base away from a terminal end of their respective polynucleotide molecules;

(3) optionally, at least one terminal base is identical in all of said polynucleotide molecules; and

(4) none of said polynucleotide molecules contains a complete sequencer motif.

45. The method of Claim 44, further comprising determining nucleotide sequences of said amplified tagged polynucleotide fragments.

46. The method of Claim 45, wherein said nucleotide sequences of said amplified tagged polynucleotide fragments are determined without polymerase chain reaction (PCR).

47. The method of Claim 45, further comprising analyzing said nucleotide sequences with a programmed computer processor to identify one or more genetic variants in said nucleotide sample of the subject.

48. The method of Claim 44, wherein said nucleic acid sample is a cell-free nucleic acid sample.

49. The method of Claim 44, wherein exposing said polynucleotide fragments of said nucleic acid sample to said plurality of polynucleotide molecules yields said tagged polynucleotide fragments with a conversion efficiency of at least 10%.

50. The method of Claim 44, wherein said subjecting comprises amplifying said tagged polynucleotide fragments from sequences corresponding to genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

51. A method, comprising:

(a) generating a plurality of sequence reads from a plurality of polynucleotide molecules, wherein said plurality of polynucleotide molecules cover genomic loci of a target genome, wherein said genomic loci correspond to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1;

(b) grouping with a computer processor said plurality of sequence reads into families, wherein each family comprises sequence reads from one of said template polynucleotides;

(c) for each of said families, merging sequence reads to generate a consensus sequence;

(d) calling said consensus sequence at a given genomic locus among said genomic loci;

(e) detecting at said given genomic locus any of:

i. genetic variants among the calls;

ii. frequency of a genetic alteration among the calls;

iii. total number of calls; and

iv. total number of alterations among the calls.

52. The method of Claim 51, wherein each family comprises sequence reads from only one of said template polynucleotides.

53. The method of Claim 51, further comprising performing (d)-(e) at an additional genomic locus among said genomic loci.

54. The method of Claim 53, further comprising determining a variation in copy number at one of said given genomic locus and additional genomic locus based on counts at said given genomic locus and additional genomic locus.

55. The method of Claim 51, wherein said grouping comprises classifying said plurality of sequence reads into families by identifying (i) distinct molecular barcodes coupled to said plurality of polynucleotide molecules and (ii) similarities between said plurality of sequence reads, wherein each family includes a plurality of nucleic acid sequences that are associated with a distinct combination of molecular barcodes and similar or identical sequence reads.

56. The method of Claim 51, wherein said consensus sequence is generated by evaluating a quantitative measure or a statistical significance level for each of said sequence reads.

57. The system of Claim 51, wherein said plurality of genes includes at least 10 of said plurality of genes selected from said group.

58. A method, comprising:

(a) providing template polynucleotide molecules and a set of library adaptors in a single reaction vessel, wherein said library adaptors are polynucleotide molecules that have different molecular barcodes, and wherein none of said library adaptors contains a complete sequencer motif;

(b) in said single reaction vessel, coupling said library adaptors to said template polynucleotide molecules at an efficiency of at least 10%, thereby tagging each template polynucleotide with a tagging combination that is among a plurality of different tagging combinations, to produce tagged polynucleotide molecules;

(c) subjecting said tagged polynucleotide molecules to an amplification reaction under conditions that yield amplified polynucleotide molecules as amplification products of said tagged polynucleotide molecules; and

(d) sequencing said amplified polynucleotide molecules.

59. The method of Claim 58, wherein said library adaptors are identical but for said molecular barcodes.

60. The method of Claim 58, wherein each of said library adaptors has a double stranded portion and at least one single-stranded portion, and wherein said single-stranded portion has a partial sequencer motif.

61. The method of Claim 58, wherein said library adaptors couple to both ends of said template polynucleotide molecules.

62. The method of Claim 58, wherein said efficiency is at least 30%.

63. The method of Claim 58, further comprising identifying genetic variants upon sequencing said amplified polynucleotide molecules.

64. The method of Claim 58, wherein said sequencing comprises (i) subjecting said amplified polynucleotide molecules to an additional amplification reaction under conditions that yield additional amplified polynucleotide molecules as amplification products of said amplified polynucleotide molecules, and (ii) sequencing said additional amplified polynucleotide molecules.

65. The method of Claim 64, wherein said additional amplification is performed in the presence of sequencing adaptors.

66. The method of Claim 58, wherein (b) and (c) are performed without aliquoting said tagged polynucleotide molecules.

67. A system for analyzing a target nucleic acid molecule of a subject, comprising:

a communication interface that receives nucleic acid sequence reads for a plurality of polynucleotide molecules that cover genomic loci of a target genome;

computer memory that stores said nucleic acid sequence reads for said plurality of polynucleotide molecules received by said communication interface; and

a computer processor operatively coupled to said communication interface and said memory and programmed to (i) group said plurality of sequence reads into families, wherein each family comprises sequence reads from one of said template polynucleotides, (ii) for each of said families, merge sequence reads to generate a consensus sequence, (iii) call said consensus sequence at a given genomic locus among said genomic loci, and (iv) detect at said given genomic locus any of genetic variants among the calls, frequency of a genetic alteration among the calls, total number of calls; and total number of alterations among the calls,

wherein said genomic loci correspond to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

68. A set of oligonucleotide molecules that selectively hybridize to at least 5 genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

69. The set of Claim 68, wherein said oligonucleotide molecules are from 10-200 bases in length.

70. The kit of Claim 68, wherein said oligonucleotide molecules selectively hybridize to exon regions of said at least 5 genes.

71. The kit of Claim 70, wherein said oligonucleotide molecules selectively hybridize to at least 30 exons in said at least 5 genes.

72. The kit of Claim 71, wherein multiple oligonucleotide molecules selectively hybridize to each of said at least 30 exons.

73. The kit of Claim 72, wherein said oligonucleotide molecules that hybridize to each exon have sequences that overlaps with at least 1 other oligonucleotide molecule.

74. A kit, comprising:

a first container containing a plurality of library adaptors each having a different molecular barcode; and

a second container containing a plurality of sequencing adaptors, each sequencing adaptor comprising at least a portion of a sequencer motif and optionally a sample barcode.

75. The kit of Claim 74, wherein said sequencing adaptor comprises said sample barcode.

76. A method for detecting sequence variants in a cell free DNA sample, comprising: detecting rare DNA at a concentration less than 1% with a specificity that is greater than 99.9%.

77. A method, comprising:

(a) providing a sample comprising a set of double-stranded polynucleotide molecules, each double-stranded polynucleotide molecule including first and second complementary strands;

(b) tagging said double-stranded polynucleotide molecules with a set of duplex tags, wherein each duplex tag differently tags said first and second complementary strands of a double-stranded polynucleotide molecule in said set;

(c) sequencing at least some of said tagged strands to produce a set of sequence reads;

(d) reducing and/or tracking redundancy in said set of sequence reads;

(e) sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in said set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among said sequence reads in said set of sequence reads;

(f) determining quantitative measures of (i) said paired reads and (ii) said unpaired reads that map to each of one or more genetic loci; and

(g) estimating with a programmed computer processor a quantitative measure of total double-stranded polynucleotide molecules in said set that map to each of said one or more genetic loci based on said quantitative measure of paired reads and unpaired reads mapping to each locus.

78. The method of Claim 77, further comprising (h) detecting copy number variation in said sample by determining a normalized total quantitative measure determined in step (g) at each of said one or more genetic loci and determining copy number variation based on the normalized measure.

79. The method of Claim 77, wherein said sample comprises double-stranded polynucleotide molecules sourced substantially from cell-free nucleic acids.

80. The method of Claim 77, wherein said duplex tags are not sequencing adaptors.

81. The method of Claim 77, wherein reducing redundancy in said set of sequence reads comprises collapsing sequence reads produced from amplified products of an original polynucleotide molecule in said sample back to said original polynucleotide molecule.

82. The method of Claim 81, further comprising determining a consensus sequence for said original polynucleotide molecule.

83. The method of Claim 82, further comprising identifying polynucleotide molecules at one or more genetic loci comprising a sequence variant.

84. The method of Claim 82, further comprising determining a quantitative measure of paired reads that map to a locus, wherein both strands of said pair comprise a sequence variant.

85. The method of Claim 84, further comprising determining a quantitative measure of paired molecules in which only one member of said pair bears a sequence variant and/or determining a quantitative measure of unpaired molecules bearing a sequence variant.

86. A method, comprising:

(a) from a sequencer, receiving into memory a set of sequence reads of polynucleotides tagged with duplex tags;

(b) reducing and/or tracking redundancy in said set of sequence reads;

(c) sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in said set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among said sequence reads in said set of sequence reads;

(d) determining quantitative measures of (i) said paired reads and (ii) said unpaired reads that map to each of one or more genetic loci; and

(e) estimating a quantitative measure of total double-stranded polynucleotide molecules in said set that map to each of said one or more genetic loci based on said quantitative measure of paired reads and unpaired reads mapping to each locus.

87. A method, comprising:

(a) providing a sample comprising a set of double-stranded polynucleotide molecules, each double-stranded polynucleotide molecule including first and second complementary strands;

(b) tagging said double-stranded polynucleotide molecules with a set of duplex tags, wherein each duplex tag differently tags said first and second complementary strands of a double-stranded polynucleotide molecule in said set;

(c) sequencing at least some of said tagged strands to produce a set of sequence reads;

(d) reducing and/or tracking redundancy in said set of sequence reads;

(e) sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in said set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among said sequence reads in said set of sequence reads; and

(f) determining quantitative measures of at least two of (i) said paired reads, (ii) said unpaired reads that map to each of one or more genetic loci, (iii) read depth of said paired reads and (iv) read depth of unpaired reads.

88. A method, comprising:

(a) tagging control parent polynucleotides with a first tag set to produce tagged control parent polynucleotides, wherein said first tag set comprises a plurality of tags, wherein each tag in said first tag set comprises a same control tag and an identifying tag, and wherein said tag set comprises a plurality of different identifying tags;

(b) tagging test parent polynucleotides with a second tag set to produce tagged test parent polynucleotides, wherein said second tag set comprises a plurality of tags, wherein each tag in said second tag set comprises a same test tag that is distinguishable from said control tag and an identifying tag, and wherein said second tag set comprises a plurality of different identifying tags;

(c) mixing tagged control parent polynucleotides with tagged test parent polynucleotides to form a pool;

(d) amplifying tagged parent polynucleotides in said pool to form a pool of amplified, tagged polynucleotides;

(e) sequencing amplified, tagged polynucleotides in said amplified pool to produce a plurality of sequence reads;

(f) grouping sequence reads into families, each family comprising sequence reads generated from a same parent polynucleotide, which grouping is optionally based on information from an identifying tag and from start/end sequences of said parent polynucleotides, and, optionally, determining a consensus sequence for each of a

plurality of parent polynucleotides from said plurality of sequence reads in a group;

(g) classifying each family or consensus sequence as a control parent polynucleotide or as a test parent polynucleotide based on having a test tag or a control tag;

(h) determining a quantitative measure of control parent polynucleotides and control test polynucleotides mapping to each of at least two genetic loci; and

(i) determining copy number variation in said test parent polynucleotides at at least one locus based on relative quantity of test parent polynucleotides and control parent polynucleotides mapping to said at least one locus.

89. A method comprising:

(a) generating a plurality of sequence reads from a plurality of template polynucleotides, each polynucleotide mapped to a genomic locus;

(b) grouping the sequence reads into families, each family comprising sequence reads generated from one of the template polynucleotides;

(c) calling a nucleotide base or sequence at the genomic locus for each of the families;

(d) detecting at the genomic locus any of:

i. genomic alterations among the calls;

ii. frequency of a genetic alteration among the calls;

iii. total number of calls;

iv. total number of alterations among the calls.

90. The method of Claim 89, wherein calling comprises any of phylogenetic analysis, voting, weighing, assigning a probability to each read at the locus in a family, and calling the nucleotide base with the highest probability.

91. The method of Claim 89, performed at two loci, comprising determining CNV at one of the loci based on counts at each of the loci.

92. A method comprising:

(a) ligating adaptors to double-stranded deoxyribonucleic acid (DNA) polynucleotides, wherein ligating is performed in a single reaction vessel, and wherein the adaptors comprise molecular barcodes, to produce a

tagged library comprising an insert from the double-stranded DNA polynucleotides, and having between 4 and 1 million different tags;

(b) generating a plurality of sequence reads for each of said double-stranded DNA polynucleotides in the tagged library;

(c) grouping sequence reads into families, each family comprising sequence reads generated from a single DNA polynucleotide among said double-stranded DNA polynucleotides, based on information in a tag and information at an end of the insert; and

(d) calling nucleotide bases at each position in the double-stranded DNA molecule based on nucleotide bases at the position in members of a family.

93. The method of Claim 93, wherein (d) comprises calling a plurality of sequential bases from at least a subset of said sequence reads to identify single nucleotide variations (SNV) in the double-stranded DNA molecule.