

METHODS FOR THE DETECTION OF GENOMIC COPY CHANGES IN DNA SAMPLES

REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to U.S. Provisional Patent Application No. 62/379,593, filed August 25, 2016, and U.S. Provisional Patent Application No. 62/481,538, filed April 4, 2017, each of which are incorporated herein by reference in their entireties.

STATEMENT REGARDING SEQUENCE LISTING

[0002] The sequence listing associated with this application is provided in text format in lieu of a paper copy, and is hereby incorporated by reference into the specification. The name of the text file containing the sequence listing is CLFK_005_02WO_ST25. The text file is 2,238 KB, was created on August 24, 2017, and is being submitted electronically via EFS-Web.

TECHNICAL FIELD

[0003] The invention relates generally to compositions and methods for the quantitative genetic analysis of biological samples, *e.g.*, direct tissue biopsies or peripheral blood. In particular, the present invention relates to methods for detection of target-specific copy number change, as well as genetic characterization and analysis, of biological samples.

BACKGROUND

[0004] It is becoming increasingly clear that most, if not all, of the most common human cancers are diseases of the human genome. It is thought that somatic mutations accumulate during an individual's lifetime, some of which increase the probability that the cell in which they are harbored can develop into a tumor. With just the wrong combination of accumulated mutational events, a precancerous growth loses constraints that keep uncontrolled proliferation in check and the resulting cell mass becomes a cancer. The constellations of mutations that are necessary and sufficient to cause cancer are often collectively referred to as "driver mutations." One of the themes that have emerged from recent and intensive molecular analysis is that cancer, once thought of as a single, tissue-specific disease, is in fact a group of related diseases, each with a unique molecular

pathology. The human genome project laid the groundwork for genome-wide analysis of cancers.

[0005] Changes in gene copy number are a fundamental driver of biological diversity. In the context of evolution, duplication of genes and divergence of function is a well-recognized driver of species diversity. In the context of human disease, gene loss and gene amplification within somatic cells are hallmarks of diseased tissues such as cancer. Certain therapeutic agents act specifically on cells with these genomic gain and/or loss mutations, however, the identification of these copy number variations is difficult because often such mutations are only present within the DNA of diseased or cancerous cells and are not found in other cells of the body. While the diseased tissue or cells is the major source of the mutated DNA, acquiring DNA through a biopsy is invasive, risky and often not possible. The observation that dying tumor or cancer cells release small pieces of their DNA into the bloodstream, termed cell free DNA or circulating DNA has allowed for the development of genetic tests that can be performed with less invasive techniques, such as a blood sample. However, only small amounts of DNA can be obtained from isolating cell free DNA from a sample, and only a portion of the total DNA will carry the mutation associated with the disease. For example, in the context of cancer genomics, diagnostically significant tumor mutations are often only found at minor allele frequencies that are significantly less than 50%. This is in contrast to conventional SNP genotyping where allele frequencies are generally ~100%, 50% or 0%.

[0006] Thus there is a need for genomic techniques capable of detecting genetic copy number changes in specific target loci.

SUMMARY

[0007] Methods of detecting rare mutations in cfDNA have been previously described in International PCT Publication No. WO 2016/028316. However, these techniques still lack the requisite sensitivity to detect the rarest copy number losses at very minor allele frequencies. Provided herein are compositions and methods for detection of target-specific copy number change that are applicable to several sample types, including direct tissue biopsies, peripheral blood, and in particular cfDNA. The compositions and methods described herein are sensitive enough to detect changes in copy number that are present only a tiny fraction of the total DNA.

[0008] The present disclosure includes, *inter alia*, compositions and methods that are useful for the detection of a mutational change, SNP, translocation, inversion, deletion, change in copy number, or other genetic variation within a sample of cellular genomic DNA (*e.g.*, from a tissue biopsy sample) or cfDNA (*e.g.*, from a blood sample). In particular, the compositions and methods of the present disclosure provide an extremely high level of resolution that is particularly useful in detecting copy number variations in a small fraction of the total cfDNA from a biological sample (*e.g.*, blood).

[0009] Particular examples are drawn to a method for performing a genetic analysis on a DNA target region from a test sample comprising: (a) generating a genomic DNA library comprising a plurality of DNA library fragments, wherein each of the DNA library fragments comprises a genomic DNA fragment from the test sample and an adaptor; (b) contacting the genomic DNA library with a plurality of capture probes that specifically bind to a DNA target region, thereby forming complexes between the capture probes and DNA library fragments comprising the DNA target region; and (c) performing a quantitative genetic analysis of the genomic DNA fragments comprising the DNA target region; wherein the adaptor is a DNA polynucleotide that comprises: an amplification region, a sample tag region, and an anchor region; wherein the amplification region comprises a polynucleotide sequence capable of serving as a primer recognition site for PCR amplification; wherein the sample tag comprises a polynucleotide sequence that encodes an identity of the unique library DNA fragment and encodes an identity of the test sample; wherein the anchor region comprises a polynucleotide sequence that encodes the identity of the test sample and wherein the anchor region is capable of attaching to the genomic DNA fragment; and wherein the genetic analysis is performed to detect a genetic change indicative of a disease state.

[0009a] In one aspect, the present disclosure refers to a method for performing a genetic analysis on a DNA target region from a test sample comprising a plurality of genomic DNA fragments, the method comprising: (a) generating a genomic DNA library by contacting the test sample with a set of adaptors, wherein each adaptor of the set of adaptors comprises a sample tag region selected from a pool of unique sample tag regions, wherein the pool is selected from a plurality of pools, and wherein the selected pool is unique to the test sample; wherein the genomic DNA library comprises a plurality of library DNA fragments and at least two unique sample tag regions; wherein each of the library DNA fragments comprises a genomic DNA fragment from the test sample attached to an adaptor; (b) contacting the genomic DNA library with a plurality of capture probes that specifically bind to the DNA target region, thereby forming complexes between the capture probes and library DNA fragments comprising the DNA target region; and (c) performing a quantitative genetic analysis of the genomic DNA fragments comprising the DNA target region; wherein the adaptor is a DNA polynucleotide that comprises: (i) an amplification region, (ii) a sample tag region, and (iii) an anchor region; wherein the amplification region comprises a polynucleotide sequence capable of serving as a primer recognition site for PCR amplification; wherein the sample tag region identifies the test sample; and wherein the anchor region comprises a polynucleotide sequence that is capable of attaching to the genomic DNA fragment.

[0010] In some examples, the genetic change indicative of a disease state is selected from a single nucleotide variant (SNV), an insertion less than 40 nucleotides in length, a deletion of a DNA region less than 40 nucleotides in length, and/or a change in copy number. In particular examples, the genetic change indicative of a disease state is a change in copy number. In some examples, the test sample is a tissue biopsy. In various examples, the tissue biopsy is taken from a tumor or a tissue suspected of being a tumor. In certain examples, the genomic DNA is cell free DNA (cfDNA) or cellular DNA. In particular examples, the genomic DNA is cfDNA is isolated from the test sample; and wherein the test sample is a biological sample selected from the group consisting of: amniotic

fluid, blood, plasma, serum, semen, lymphatic fluid, cerebral spinal fluid, ocular fluid, urine, saliva, stool, mucous, and sweat.

[0011] In certain examples, the genomic DNA fragments are obtained the steps comprising; (i) isolating cellular DNA from the test sample; and (ii) fragmenting the cellular DNA to obtain the genomic DNA fragments. In particular examples, step (ii) is performed by contacting the cellular DNA with at least one digestion enzyme. In some examples, step (ii) is performed by applying mechanical stress to the cellular DNA. In certain examples, the mechanical stress is applied by sonicating the cellular DNA.

[0012] In particular examples, the sample tag further comprises a unique molecule identifier (UMI) that facilitates the identification of the unique genomic DNA fragment.

[0013] In some examples, the amplification region is between 10 and 50 nucleotides in length. In particular examples, the amplification region is between 20 and 30 nucleotides in length. In certain examples, the amplification region is 25 nucleotides in length.

[0014] In some examples, the sample tag is between 5 and 50 nucleotides in length. In particular examples, the sample tag is between 5 and 15 nucleotides in length. In certain examples, the sample tag is 8 nucleotides in length. In some examples, the UMI multiplier is adjacent to or contained within the sample tag region.

[0015] In certain examples, the UMI multiplier is between 1 and 5 nucleotides in length. In particular examples, the UMI multiplier is 3 nucleotides in length, and comprises one of 64 possible nucleotide sequences.

[0016] In some examples, the anchor region is between 1 and 50 nucleotides in length. In particular examples, the anchor region is between 5 and 25 nucleotides in length. In certain examples, the anchor region is 10 nucleotides in length.

[0017] Particular examples of the present disclosure are drawn to methods where the step of (a) generating a genomic DNA library comprising a plurality of DNA library fragments, comprises attaching the genomic DNA fragments to a plurality of adaptors. In certain examples, the genomic DNA fragments are end repaired prior to attaching the genomic DNA fragments with a plurality of adaptors. In particular examples, the amplification regions of each adaptor of the plurality of adaptors comprises an identical nucleotide sequence.

[0018] In certain examples, the sample tag region of each adaptor of the plurality of adaptors comprise one of between 2 and 1,000 nucleotide sequences. In particular examples, the sample tag region of each adaptor of the plurality of adaptors comprise one of between 50 and 500 nucleotide sequences. In various examples, the sample tag region of each adaptor of the plurality of adaptors comprises one of between 100 and 400 nucleotide sequences. In some examples, the sample tag region of each adaptor of the plurality of adaptors comprises one of between 200 and 300 nucleotide sequences. In certain examples, the sample tag region of each adaptor of the plurality of adaptors is 8 nucleotides in length. In some examples, each sequence of the nucleotide sequences are discrete from any other sequence of the 240 nucleotide sequences by Hamming distance of at least two.

[0019] In particular examples, each of the plurality of adaptors comprises a UMI multiplier that is adjacent to or contained within the sample tag region. In some examples, each of the plurality of adaptors comprises a UMI multiplier that is adjacent to the sample tag region. In certain examples, the UMI multiplier of each adaptor of the plurality of adaptors is between 1 and 5 nucleotides in length. In some examples, the UMI multiplier of each adaptor of the plurality of adaptors is three nucleotides in length.

[0020] In particular examples, the anchor tag region of each adaptor of the plurality of adaptors comprises one of four nucleotide sequences, and each sample region of a given sequence is paired to only one of the four anchor regions of a given sequence.

[0021] In some examples, the amplification regions of each adaptor of the plurality of adaptors comprises an identical nucleotide sequence; the sample tag region of each adaptor of the plurality of adaptors is 8 nucleotides in length; the nucleotide sequence of each sample tag is discrete from any other nucleotide sequence of the sample tags of the plurality of adaptors by Hamming distance of at least two; each of the plurality of adaptors comprises a UMI multiplier that is adjacent to or contained within the sample tag region; the UMI multiplier of each adaptor of the plurality of adaptors is three nucleotides in length; and the UMI multiplier of each of the possible nucleotide sequences is paired to each sample tag region of the plurality of adaptors; the anchor tag region of each adaptor of the plurality of adaptors comprises one of four nucleotide sequences; and each sample region of a given sequence is paired to only one of the four anchor regions of a given sequence.

[0022] Particular examples of the present disclosure are drawn to a method where the step of attaching the genomic DNA fragments with a plurality of adaptors

comprises: (i) attaching an oligonucleotide comprising least a portion of an anchor region to each genomic DNA fragment, wherein the oligonucleotide comprising least a portion of an anchor region is a DNA duplex comprising a 5' phosphorylated attachment strand duplexed with a partner strand, wherein the partner strand is blocked from attachment by chemical modification at its 3' end, and wherein the attachment strand is attached to the genomic DNA fragment; (ii) contacting the genomic DNA fragments attached to the oligonucleotides comprising at least a portion of the anchor region with DNA oligonucleotides encoding full length adaptor sequences for each adaptor nucleotide sequence of the plurality of adaptors; and (iii) contacting the genomic DNA fragments and the DNA oligonucleotides encoding the full length adaptor sequence with T4 polynucleotide kinase, Taq DNA ligase and full-length Bst polymerase under conditions suitable for DNA ligation; thereby attaching the plurality of adaptors to the genomic DNA fragments. In some examples, the genomic DNA fragments are cfDNA. In certain examples, the DNA target region is analyzed for a change in copy number.

[0023] In particular examples, step (c) performing a quantitative genetic analysis of the genomic DNA fragments comprising the DNA target region comprises purification of the complexes formed between the capture probes and DNA library fragments comprising the DNA target region. In certain examples, step (c) comprises purification of the complexes formed between the capture probes and DNA library fragments comprising the DNA target region, performing primer extension and/or amplification of the DNA library fragments comprising the region of interest from the genomic DNA library. In some examples, step (c) comprises purification of the complexes formed between the capture probes and DNA library fragments comprising the DNA target region, performing primer extension and amplification of the DNA library fragments comprising the region of interest from the genomic DNA library. In certain examples, step (c) comprises DNA sequencing of the DNA library fragments comprising the DNA target region to generate a plurality of sequencing reads.

[0024] In some examples, the present disclosure is drawn to a method wherein the genomic analysis comprises determining a change of copy number in a DNA region of interest, and wherein step (c), performing a quantitative genetic analysis of the genomic DNA fragments comprising the DNA target region, comprises determining a copy number of the region of interest present in the genomic DNA library derived from the test sample, and comparing it to a copy number of the region of interest present in the genomic

DNA library derived from a reference sample, wherein the reference sample comprises a known copy number of the DNA target region.

[0025] In some examples, determining the copy number in the region of interest comprises DNA sequencing of the DNA library fragments comprising the DNA target region to generate a plurality of sequencing reads, wherein each sequencing read comprises a unique molecular identification element (UMIE). In some examples, the UMIE comprises sequencing information from the adaptor and at least a portion of the genomic DNA sequence. In some examples, sequencing reads comprising identical UMIEs are identified as a unique genomic sequence (UGS).

[0026] In some examples, methods of determining the copy number further comprise determining a raw genomic depth (RGD) for each of the capture probes contacted with the genomic DNA library. In some examples, determining the RGD comprises determining the average number of UGSs associated with each capture probe sequence within a group of sample replicates. In some examples, capture probes associated with a highly variable number of UGSs are identified as noisy probes and are removed from further calculations. In some examples, determining the RGD further comprises calculating an RGD for a sample, comprising calculating a numerical average of all RGDs for all capture probes in the sample. In some examples, the RGD values for noisy probes are not included in calculating an RGD for a sample.

[0027] In some examples, the RGDs for the capture probes are normalized across all samples in an experimental group by converting the RGD for each capture probe into a probe-specific, normalized read count comprising (i) multiplying each capture probe RGD in a sample by a normalization constant, wherein the normalization constant comprises any real number; and (ii) dividing the product of (i) by the RGD calculated for the corresponding sample; or (iii) dividing the product of (i) by an average RGD calculated from a subset of probes. In some examples, the subset of probes is a set of control probes.

[0028] In some examples, the probe-specific, normalized read counts are converted in to a copy number value comprising (i) multiplying the probe-specific, normalized read counts of probes directed to autosomal and/or X-linked regions by 2 in samples derived from females; (ii) multiplying the probe-specific, normalized read counts of probes directed to Y-linked and/or X-linked regions by 1 in samples derived from males; (iii) averaging the products of (i) and/or (ii) across all samples in an experiment; and (iv) dividing

the product of (i) and/or (ii) by the average of (iii). In some examples, the approximate copy number values for all probes that target a specific gene are averaged.

[0029] In some examples, the present disclosure is drawn to a method for highly sensitive detection of copy number gain and copy number loss comprising (i) determining an RGD for a capture probe; (ii) normalizing the RGD for the capture probe across all samples in an experimental group by converting the RGD for the capture probe into a probe-specific, normalized read count; (iii) calculating an approximate copy number value for each probe-specific, normalized read count; and (iv) averaging the approximate copy number values for all probes that target a specific gene.

[0030] In some examples, the present disclosure is drawn to a method for measuring chromosome stability comprising (i) designing and validating a set of one or more chromosomal stability probes, wherein the chromosomal stability probes are uniformly distributed across human chromosomes; (ii) performing targeted sequencing on patient samples using the one or more chromosomal stability probes; (iii) determining an approximate copy number value for each chromosomal probe; (iv) determining a genomic phenotype of a patient sample, wherein fluctuations in the copy number values for one or more chromosomal probes in the patient sample indicate genomic instability.

[0031] In some examples, the present disclosure is drawn to a method of treating a cancer in a subject in need thereof, wherein the subject has been identified as having a destabilized genome according to the method claim 62, wherein the method of treating the cancer comprises administering a pharmaceutically effective amount of a PARP inhibitor.

[0032] In some examples, the present disclosure is drawn to a method wherein the genomic analysis comprises determining a change of copy number in a DNA region of interest, and wherein step (c), performing a quantitative genetic analysis of the genomic DNA fragments comprising the DNA target region, comprises determining a copy number of the region of interest present in the genomic DNA library derived from the test sample, and comparing it to a copy number of the region of interest present in the genomic DNA library derived from a reference sample, wherein the reference sample comprises a known copy number of the DNA target region. In some examples, the region of interest is a gene or a portion of the gene. In particular examples, the gene is associated with a disease. In certain examples, the disease is a cancer. In various examples, the gene is BRCA2, ATM, BRCA1, BRIP1, CHEK2, FANCA, HDAC2, and/or PALB2.

[0033] Particular examples are drawn to a genomic DNA library comprising a plurality of DNA library fragments, wherein each of the DNA library fragments comprises an adaptor and a genomic DNA fragment; wherein the adaptor is a DNA polynucleotide that comprises: an amplification region, a sample tag region, and an anchor region; wherein the amplification region comprises a polynucleotide sequence capable of serving as a primer recognition site for PCR amplification; wherein the sample tag comprises a polynucleotide sequence that encodes an identity of the unique library DNA fragment and encodes an identity of the test sample; and wherein the anchor region comprises a polynucleotide sequence that encodes the identity of the test sample, and wherein the anchor region is capable of attaching to the genomic DNA fragment. In some examples, the sample tag further comprises a unique molecule identifier (UMI), wherein the UMI facilitates the identification of the unique genomic DNA fragment. In particular examples, the amplification region is between 10 and 50 nucleotides in length. In particular examples, the amplification region is 25 nucleotides in length. In particular examples, the sample tag is between 5 and 50 nucleotides in length. In certain examples, the sample tag is 8 nucleotides in length. In some examples, the UMI multiplier is adjacent to or contained within the sample tag region. In particular examples, the UMI multiplier is between 1 and 5 nucleotides in length. In certain examples, the anchor region is between 1 and 50 nucleotides in length. In some examples, the anchor region is 10 nucleotides in length. In particular examples, the amplification regions of each adaptor of the plurality of adaptors comprises an identical nucleotide sequence. In some examples, each nucleotide sequence of the sample tags are discrete from any other sequence of the nucleotide sequences of the sample by Hamming distance of at least two. In certain examples, each of the plurality of adaptors comprises a UMI multiplier that is adjacent to or contained within the sample tag region. In particular examples, each of the plurality of adaptors comprises a UMI multiplier that is adjacent to the sample tag region. In some examples, the anchor tag region of each adaptor of the plurality of adaptors comprises one of four nucleotide sequences, and wherein each sample region of a given sequence is paired to only one of the four anchor regions of a given sequence. In some examples, the genomic DNA fragment is cfDNA.

[0034] In certain examples, the amplification regions of each adaptor of the plurality of adaptors comprises an identical nucleotide sequence; the sample tag region of each adaptor of the plurality of adaptors is 8 nucleotides in length, the sample tag region of

each adaptor of the plurality of adaptors comprises a nucleotide sequence that is discrete from any other nucleotide sequence of the sample tags of the plurality of adaptors by Hamming distance of at least two, the each of the plurality of adaptors comprises a UMI multiplier that is adjacent to or contained within the sample tag region, the UMI multiplier of each adaptor of the plurality of adaptors is three nucleotides in length, and the UMI multiplier of each of the possible nucleotide sequences is paired to each of the sample tag regions of the plurality of adaptors, the anchor tag region of each adaptor of the plurality of adaptors comprises one of four nucleotide sequences, and each sample region of a given sequence is paired to only one of the four anchor regions of a given sequence. In some examples, the genomic DNA fragment is cfDNA.

[0035] Certain examples are drawn to a plurality of genomic DNA libraries, comprising more than one genomic library described herein. In some examples, the nucleic acid sequences of the sample tag regions of a genomic DNA library belonging to the plurality of genomic DNA libraries are different from the nucleic acid sequences of the sample tag regions of other genomic DNA libraries belonging to the plurality of genomic DNA libraries. In particular examples, the nucleic acid sequences of the amplification regions of a genomic DNA library belonging to the plurality of genomic DNA libraries are identical to the nucleic acid sequences of the amplification regions of other genomic DNA libraries belonging to the plurality of genomic DNA libraries.

[0036] Certain examples are drawn to a method for genetic analysis of a DNA target region of cell free DNA (cfDNA) comprising: (a) generating a DNA library as described herein; (b) contacting the cfDNA library with a plurality of capture probes that specifically bind to a DNA target region, thereby forming complexes between the capture probes and DNA library fragments comprising the DNA target region; and (c) performing a quantitative genetic analysis of the cfDNA fragments comprising the DNA target region; thereby performing genetic analysis of the DNA target region.

[0037] Certain examples are directed to a method of predicting, diagnosing, or monitoring a genetic disease in a subject comprising: (a) obtaining a test sample from the subject; (b) isolating genomic DNA from the test sample; (c) generating a DNA library comprising a plurality of DNA library fragments, wherein each of the DNA library fragments comprises a genomic DNA fragment from the test sample and an adaptor; (d) contacting the cfDNA library with a plurality of capture probes that specifically bind to a DNA target region, thereby forming complexes between the capture probes and DNA library

fragments comprising the DNA target region; and (e) performing a quantitative genetic analysis of one or more target genetic loci associated with the genetic disease in the cfDNA clone library, wherein the identification or detection of one or more genetic lesions in the one or more target genetic loci is prognostic for, diagnostic of, or monitors the progression of the genetic disease. In particular examples, the quantitative genetic analysis comprises DNA sequencing to generate a plurality of sequencing reads.

[0038] Particular examples are drawn to a set of adaptors that encode an identify of a unique genomic DNA fragment and an identity of a test sample, for use in generating a genomic DNA library, wherein each adaptor in said set of adapters is a DNA polynucleotide that comprises: an amplification region, a sample tag region, and an anchor region; wherein the amplification region comprises a polynucleotide sequence capable of serving as a primer recognition site for PCR amplification; wherein the sample tag comprises a polynucleotide sequence that encodes the identity of the unique library DNA fragment and encodes the identity of the test sample; and wherein the anchor region comprises a polynucleotide sequence that encodes the identity of the test sample, and wherein the anchor region is capable of attaching to the genomic DNA fragment. In some examples, the sample tag further comprises a unique molecule identifier (UMI), wherein the UMI facilitates the identification of the unique genomic DNA fragment. In various examples, the amplification region is between 10 and 50 nucleotides in length. In certain examples, the amplification region is 25 nucleotides in length. In particular examples, the sample tag is between 5 and 50 nucleotides in length. In some examples, the sample tag is 8 nucleotides in length. In particular examples, the UMI multiplier is adjacent to or contained within the sample tag region. In some examples, the UMI multiplier is between 1 and 5 nucleotides in length. In particular examples, the anchor region is between 1 and 50 nucleotides in length. In some examples, the anchor region is 10 nucleotides in length. In certain examples, the amplification regions of each adaptor of the plurality of adaptors comprises an identical nucleotide sequence.

[0039] In some examples, each nucleotide sequence of the sample tags is discrete from any other nucleotide sequence of the sample tags of the set of adaptors by Hamming distance of at least two. In various examples, each of the plurality of adaptors comprises a UMI multiplier that is adjacent to or contained within the sample tag region. In particular examples, each of the plurality of adaptors comprises a UMI multiplier that is adjacent to the sample tag region.

[0040] In some examples, the anchor tag region of each adaptor of the plurality of adaptors comprises one of four nucleotide sequences, and wherein each sample region of a given sequence is paired to only one of the four anchor regions of a given sequence. The set of adaptors claim 75, wherein the amplification regions of each adaptor of the plurality of adaptors comprises an identical nucleotide sequence; wherein the sample tag region of each adaptor is 8 nucleotides in length, wherein each nucleotide sequence of the sample tags is discrete from any other nucleotide sequence of the sample tags of the set of adaptors by Hamming distance of at least two, wherein each of the plurality of adaptors comprises a UMI multiplier that is adjacent to or contained within the sample tag region, wherein the UMI multiplier of each adaptor of the plurality of adaptors is three nucleotides in length, wherein the UMI multiplier comprises one of 64 possible nucleotide sequences, and wherein the UMI multiplier of each of the 64 possible nucleotide sequences is paired to each of the sample tag region of the plurality of adaptors, wherein the anchor tag region of each adaptor of the plurality of adaptors comprises one of four nucleotide sequences, and wherein each sample region of a given sequence is paired to only one of the four anchor regions of a given sequence.

BRIEF DESCRIPTION OF THE SEVERAL VIEWS OF THE DRAWINGS

[0041] FIG. 1 shows the framework of the copy number loss (CNL) assay. Each gene (rows) exhibits a characteristic unique read value that is represented here by a shade. Each sample (columns) is interrogated across the same panel of genes.

[0042] FIG. 2 shows a diagram illustrating the drivers of the CNL assay signal.

[0043] FIG. 3 shows a diagram illustrating steps of an illustrative CNL assay performed on cell free DNA (cfDNA).

[0044] FIG. 4A – 4E shows diagrams of an illustrative first generation adaptor (FIG. 4A and 4B) and an adaptor of the present invention (FIGs. 4C-4E). FIG. 4A shows the first generation adaptor design. FIG. 4B shows that in the first generation adaptors, there were a collection of 249 possible sequence tags, each 5 nucleotides (nt) in length that attached to a single anchor sequence. FIG. 4C shows a diagram of a second generation adaptor. FIG. 4D shows an illustrative set of adaptors that are applied to a single sample that consists of four sets of 8mer tag sequences with each set having 60 members. Each set of 60

tags is specific to one of four anchor sequences. FIG. 4E shows an illustrative DNA sequence of a 47 nt adaptor.

[0045] FIG. 5A – FIG. 5B shows a diagram illustrating that shifting the position of the UMI multiplier within the sample tag can increase the number of unique sample tags.

[0046] FIG. 6A and B shows a diagram illustrating the process of constructing genomic libraries for a CNL assay. FIG. 6A shows the step where the 10 nt anchor sequence is attached to the 3' ends of genomic fragments. FIG. 6B shows the step where the full length genomic adaptors are annealed to the initial anchor sequence.

[0047] FIG. 7 shows DNA inputs into CNL libraries. Agarose gel images are shown with the sizes of markers (bp) indicated at left.

[0048] FIG. 8A – FIG. 8C shows conventional box-and-whiskers plots of measured gene copies across eight samples as determined by CNL analysis.

[0049] FIG. 9A – FIG. 9B shows Log₁₀ P-value plots that quantify significant deviation-from-normal in CNL measurements for fragmented genomic samples. The SNP percentages at the top show the minor allele frequencies of rare, heterozygous SNPs that are present in the Δ ATM and Δ BRCA2 samples.

[0050] FIG. 10A – FIG. 10B shows Log₁₀ P-value plots that quantify significant deviation-from-normal in CNL measurements for cfDNA samples spiked with fragmented genomic DNA. The SNP percentages at the top show the minor allele frequencies of rare, heterozygous SNPs that are present in the Δ ATM and Δ BRCA2 samples.

[0051] FIG. 11A – 11D illustrate the targeted hybrid capture platform. FIG. 11A shows conversion of cfDNA to a genomic library by the addition of adaptor sequences that provide universal, single-primer PCR amplification sequences, sample multiplexing tags, and unique molecular identifiers to every genomic clone. FIG. 11B shows denatured amplified genomic hybridized with target specific capture probes and primer extension. FIG. 11C shows a schematic of asymmetric paired-end sequencing. FIG. 11D shows mapping statistics for 377,711,020 Illumina NextSeq reads from a typical targeted capture sequence run. 98.5% of reads map to their intended targets. Following de-duplication, 20.40% of reads (77,053,048) are derived from unique genomic clones.

[0052] FIG. 12A – FIG. 12H shows sequences of adaptor oligonucleotides from Pools 1 – 3.

[0053] FIG. 13A – FIG. 13H shows sequences of adaptor oligonucleotides from Pools 4 – 6.

[0054] FIG. 14A – FIG. 14I shows sequences of adaptor oligonucleotides from Pools 7 – 9.

[0055] FIG. 15A – FIG. 15H shows sequences of adaptor oligonucleotides from Pools 10 – 12.

[0056] FIG. 16A – FIG. 16H shows sequences of adaptor oligonucleotides from Pools 13 – 15.

[0057] FIG. 17A – FIG. 18H shows sequences of adaptor oligonucleotides from Pools 16 – 18.

[0058] FIG. 18A – FIG. 18H shows sequences of adaptor oligonucleotides from Pools 19 – 21.

[0059] FIG. 19A – FIG. 19H shows sequences of adaptor oligonucleotides from Pools 22 – 24.

[0060] FIG. 20A – FIG. 20H shows sequences of adaptor oligonucleotides from Pools 25 – 27.

[0061] FIG. 21A – FIG. 21H shows sequences of adaptor oligonucleotides from Pools 28 – 30.

[0062] FIG. 22A – FIG. 22H shows sequences of adaptor oligonucleotides from Pools 31 – 32.

[0063] FIG. 23A – 23C shows targeted sequencing of the TP53 gene. FIG. 23A illustrates BedFile display of capture probes. FIG. 23B illustrates coverage depth at each base position on a scale of 0 to 8000 unique reads. FIG. 23C illustrates a UCSC gene model display of known TP53 splice variants. The thicker rectangular regions represent the amino acid coding regions for the TP53-encoded protein.

[0064] FIG. 24A – 24C illustrate raw and normalized unique read density for a single probe, TP53r10_1, across 16 samples. FIG. 24A illustrates the number of raw unique reads capture by probe TP53r10_1 for 16 independent sample after removal of redundant reads by “de-duplication.” FIG. 24B shows global average of unique reads across 2596 capture probes for all 16 samples. FIG. 24C shows normalized unique read depth across 16 samples (Calculated as: [sample *n* unique reads from probe TP53r10_1 x constant ÷ global average unique reads/probe from sample *n*]).

[0065] FIG. 25 shows general consistency of the normalized unique read counts for all 16 samples within any given TP53 probe despite significant average depth variation between probes. The normalized unique read counts for all 16 samples are shown as “pillars” of tightly spaced bar graphs; the results for all 45 probes that target TP53 are shown. Two probes that exhibit “noisy” counting behavior are highlighted with arrows. Counts from such probes often appear as outliers in subsequent copy number analysis.

[0066] FIG. 26 illustrates sample-to-sample consistency of normalized probe-by-probe unique read counts across a broad panel of 2596 probes. The scatter plots from three representative samples are shown. Each dot represents a different probe. The x-axis is the normalized average unique read depth per probe across 16 samples. The y-axis is the normalized unique read depth per probe for three different individual samples. The consistent probe-by-probe unique read counts support quantitative analysis of chromosomal copy variation.

[0067] FIG. 27A – 27C illustrate copy number analysis of cfDNA from a healthy female and male donor and from an advanced stage prostate cancer patient. FIG. 27A shows analysis of a cfDNA from a healthy female donor. The x-axis is a series of control probes that target regions from all 22 autosomal chromosomes, a series of probes that target the X-linked AR gene, and a series of probes that target the coding regions of the TP53 gene. The Y-axis shows the calculated ploidy for each probe. This approximation is calculated for each probe by normalizing the observed unique read counts to a series of control samples whose ploidy is known ($[\text{unique read count for probe_Y of sample_Z}] \times 2 \div [\text{average unique read count for probe_Y for multiple control samples}]$). FIG. 27B illustrates that the X-linked AR gene exhibits a haploid copy number in healthy males. FIG. 27C illustrates copy number analysis of cfDNA from an advanced prostate cancer patient and shows evidence of very significant aneuploidy across the control probes, amplification of the AR gene, and loss of the TP53 gene.

[0068] FIG. 28 shows whole genome aneuploidy analysis of a prostate patient cfDNA library relative to a control sample. The approximate ploidy for each of 239 control probes is shown sorted by chromosome. Patient chromosome 2 probes show consistent copy loss and the majority of chromosome 5 probes show copy gain. Significant deviation of approximate ploidy are seen for many, but not all, of the patient control probes.

[0069] FIG. 29 shows analytical validation of copy number loss detection. Genomic DNA from immortalized line NA02718 (monoallelic Δ ATM) and from NA09596

(monoallelic Δ BRCA2) were spiked into the “gold standard” genomic DNA from NA12878 at 16%, resulting in the equivalent of an 8% biallelic deletion minor allele frequency. Following targeted sequencing and CNV analysis, the probe-by-probe ploidies were averaged for the two target genes. Two unperturbed control genes, BRIP1 and HDAC2, are shown for comparison.

DETAILED DESCRIPTION

A. OVERVIEW

[0070] The present invention includes, *inter alia*, compositions and methods that are useful for the detection of a mutational change, SNP, translocation, inversion, deletion, change in copy number or other genetic variation within a sample of cellular genomic DNA (*e.g.* from a tissue biopsy sample) or cfDNA (*e.g.* from a blood sample). The compositions and methods of the current invention are particularly useful in detecting incredibly hard to detect copy number variations in cfDNA from a biological sample (*e.g.* blood) with exquisite resolution. In particular, some embodiments of the present invention are drawn to a method for the detecting copy number of a DNA target region from a test sample by generating a genomic DNA library made up of genomic DNA fragments attached to an adaptor, capturing DNA target regions with a plurality of capture probes, isolating the DNA library fragments comprising the DNA target region, and performing a quantitative genetic analysis of the DNA target region to thereby determining the copy number of the DNA target region. The adaptors described herein allow for the identification of the individual DNA fragment that is being sequenced, as well as the identity of the sample or source of the genomic DNA.

[0071] The present invention contemplates, in part, compositions and methods for detection of target-specific copy number changes that are applicable to several sample types, including but not limited to direct tissue biopsies and peripheral blood. In the context of cancer genomics, and in particular cell free DNA (cfDNA) assays for the analysis of solid tumors, the amount of tumor DNA is often a very small fraction of the overall DNA. Further, copy number loss is difficult to detect in genomic DNA assays, and in particular, genomic DNA assays where copy number change may only be present in a portion of the total genomic DNA from a sample, *e.g.*, cfDNA assays. For example, most of the cell-free DNA extracted from a cancer patient will be derived from normal sources and have a diploid copy number (except for X-linked genes in male subjects). In a cancer patient, the fraction of DNA

derived from tumors often has a low minor allele frequency, such as for example, a patient in which 2% of the circulating DNA extracted from plasma is derived from the tumor. The loss of one copy of a tumor suppressor gene (for example, BRCA1 in breast cancer) means that the minor allele frequency for the absence of detectable genomic fragments is 1%. In this scenario, a copy number loss assay engineered must be able to discriminate between 100 copies (normal) and 99 copies (heterozygous gene loss). Thus, particular embodiments contemplate that the methods and compositions of the present invention allow for the detection of copy number change with sufficient resolution to detect changes in copy number at minor allele frequencies even in the context of cfDNA.

[0072] To achieve this level of discrimination, the present invention provides novel sample adaptor designs. The adaptors of the present invention are designed to include features that are critical for successful copy number loss assay performance including (i) even performance across adaptors; (ii) a high number of unique molecule identifiers (UMIs); (iii) high efficiency attachment; and (iv) accommodation of sample multiplexing. For example, the adaptors of the present invention provide the following:

[0073] Even performance across adaptors: Bioinformatics analysis often looks at intra-sample probe performance and inter-sample probe performance. Thus, it is contemplated that any performance fluctuation between adaptor pools across samples will negatively impact the ability to detect the subtle variations required by CNL analysis. In the present invention, this evenness of performance is achieved by having multiple anchor tags that are all represented in each sample tag pool, with the fixed sample tag regions (which serve to identify both the sample and the genomic fragments) being randomly selected for each pool, and a UMI multiplier that increases the unique sample tag sequences for identifying the genomic fragments.

[0074] High number of Unique Molecule Identifiers (UMIs): While adaptors must be functionally equivalent from a molecular biology perspective, they must possess a very large number of unique sequence tags ($\geq 10,000$) that augment the identification of unique genomic fragments. In this context, by “augment,” it is meant that each genomic clone fragment has a particular pair of fragmentation sites corresponding to the position in the genomic sequence where the double-strand DNA was cleaved. This cleavage site is used to differentiate unique genomic clones since each clone is likely to possess a different cleavage site. However, in libraries that possess thousands of independent clones, uniquely derived fragments will often possess the exact same cleavage sites. Genomic clones (i.e. fragments)

sharing the same cleavage site can be classified as either unique or as redundant with respect to other clone sequences derived from the same sample. By attaching adaptors that introduce a high diversity of sequence tags, different genomic clones sharing the same cleavage site are more likely to be identified as unique. In this system, the UMI is created by a combination of the sample tag region with the UMI multiplier. The combination of the UMI and the cleavage site create a unique molecular identifier element (UMIE), which facilitates the classification of sequence reads as redundant reads or unique reads. Particular embodiments contemplate that the UMI multiplier could comprise longer or shorter sequences to increase or lower the overall UMI complexity.

[0075] High efficiency attachment: Adaptors must attach to genomic fragments with high efficiency. In most oncology applications, the quantities of available cellular DNA or cfDNA are limited and therefore conversion of these genomic fragments to genomic library clones must be highly efficient. In order to achieve this, in some aspects of the present invention, the adaptor systems described herein convert about 25% to about 50% or greater of the genomic input fragments are converted into genomic library clones.

[0076] Accommodation of sample multiplexing: In general, there must be pools of different sets of adaptors where each unique adaptor of the set is attached to a different sample. At the same time, each member of the set of adaptors must possess essentially identical behavior (from a sequence counting perspective) to all other members in a set. In order to achieve this, in some embodiments, the sample tag regions have a Hamming distance of 2 between any other possible sample tag combinations reducing the chance for a read to be spuriously assigned to the wrong sample. In some embodiments, each set of adaptors is split into pools that are paired with specific anchor regions, allowing for further reduction in the possibility of an error in sample de-multiplexing. For example, in an 8mer tag with Hamming distance of 2, the total number of possible sequences is 16,384.

[0077] In a particular embodiment, pre-specified pools of adaptor oligonucleotides are provided. Such pre-specified pools are used to represent a single sample. That is, each adapter sequence in each pool of X adapter oligonucleotides (16,384 in the example given above) is distinct from each adapter sequence in every other pool used to identify other samples. One of skill in the art will recognize the number of distinct pre-specified pools that are possible for the adapter oligonucleotides will depend on the length of the sample tag and/or the UMI multiplier.

[0078] Thus, in certain embodiments the adaptors comprise a sequence, i.e., the sample tag and adjacent and/or encompassed UMI multiplier that represents or identifies both the sample and uniquely identifies the genetic fragment. This is in stark contrast to the current systems that are used in the art that use a randomly generated tag to identify the sequence and a separate barcode or sequencer indexing to allow for multiplexing.

[0079] An illustrative embodiment for detecting target-specific copy number changes within DNA obtained from a sample is shown in FIG. 3. While FIG. 3 generates a DNA library from cfDNA, this illustrative procedure could be used with DNA from other sources, e.g., fragmented cellular DNA. As shown in FIG. 3, cfDNA is collected (top panel). Next, a genomic library is generated from cfDNA by conjugating genomic library adaptors (gray circles) of the present invention to the genomic DNA. Genomic DNA fragments are captured with capture probes (black circles) that recognize the genomic region of interest. The genomic DNA of interest is sequenced, and data analysis is performed for copy loss analysis and/or characterization of the genomic DNA of interest.

[0080] The practice of particular embodiments of the invention will employ, unless indicated specifically to the contrary, conventional methods of chemistry, biochemistry, organic chemistry, molecular biology, microbiology, recombinant DNA techniques, genetics, immunology, and cell biology that are within the skill of the art, many of which are described below for the purpose of illustration. Such techniques are explained fully in the literature. See, e.g., Sambrook, *et al.*, *Molecular Cloning: A Laboratory Manual* (3rd Edition, 2001); Sambrook, *et al.*, *Molecular Cloning: A Laboratory Manual* (2nd Edition, 1989); Maniatis *et al.*, *Molecular Cloning: A Laboratory Manual* (1982); Ausubel *et al.*, *Current Protocols in Molecular Biology* (John Wiley and Sons, updated July 2008); *Short Protocols in Molecular Biology: A Compendium of Methods from Current Protocols in Molecular Biology*, Greene Pub. Associates and Wiley-Interscience; Glover, *DNA Cloning: A Practical Approach*, vol. I & II (IRL Press, Oxford, 1985); Anand, *Techniques for the Analysis of Complex Genomes*, (Academic Press, New York, 1992); *Transcription and Translation* (B. Hames & S. Higgins, Eds., 1984); Perbal, *A Practical Guide to Molecular Cloning* (1984); and Harlow and Lane, *Antibodies*, (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1998).

B. DEFINITIONS

[0081] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by those of ordinary skill in the art to which the invention belongs. Although any methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, preferred embodiments of compositions, methods and materials are described herein. For the purposes of the present invention, the following terms are defined below.

[0082] The articles “a,” “an,” and “the” are used herein to refer to one or to more than one (*i.e.* to at least one) of the grammatical object of the article. By way of example, “an element” means one element or more than one element.

[0083] The use of the alternative (*e.g.*, “or”) should be understood to mean either one, both, or any combination thereof of the alternatives.

[0084] The term “and/or” should be understood to mean either one, or both of the alternatives.

[0085] As used herein, the term “about” or “approximately” refers to a quantity, level, value, number, frequency, percentage, dimension, size, amount, weight or length that varies by as much as 15%, 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2% or 1% to a reference quantity, level, value, number, frequency, percentage, dimension, size, amount, weight or length. In one embodiment, the term “about” or “approximately” refers a range of quantity, level, value, number, frequency, percentage, dimension, size, amount, weight or length $\pm 15\%$, $\pm 10\%$, $\pm 9\%$, $\pm 8\%$, $\pm 7\%$, $\pm 6\%$, $\pm 5\%$, $\pm 4\%$, $\pm 3\%$, $\pm 2\%$, or $\pm 1\%$ about a reference quantity, level, value, number, frequency, percentage, dimension, size, amount, weight or length.

[0086] Throughout this specification, unless the context requires otherwise, the words “comprise”, “comprises,” and “comprising” will be understood to imply the inclusion of a stated step or element or group of steps or elements but not the exclusion of any other step or element or group of steps or elements. In particular embodiments, the terms “include,” “has,” “contains,” and “comprise” are used synonymously.

[0087] By “consisting of” is meant including, and limited to, whatever follows the phrase “consisting of.” Thus, the phrase “consisting of” indicates that the listed elements are required or mandatory, and that no other elements may be present.

[0088] By “consisting essentially of” is meant including any elements listed after the phrase, and limited to other elements that do not interfere with or contribute to the

activity or action specified in the disclosure for the listed elements. Thus, the phrase “consisting essentially of” indicates that the listed elements are required or mandatory, but that no other elements are optional and may or may not be present depending upon whether or not they affect the activity or action of the listed elements.

[0089] Reference throughout this specification to “one embodiment,” “an embodiment,” “a particular embodiment,” “a related embodiment,” “a certain embodiment,” “an additional embodiment,” or “a further embodiment” or combinations thereof means that a particular feature, structure or characteristic described in connection with the embodiment is included in at least one embodiment of the present invention. Thus, the appearances of the foregoing phrases in various places throughout this specification are not necessarily all referring to the same embodiment. Furthermore, the particular features, structures, or characteristics may be combined in any suitable manner in one or more embodiments.

[0090] As used herein, the term “isolated” means material that is substantially or essentially free from components that normally accompany it in its native state. In particular embodiments, the term “obtained” or “derived” is used synonymously with isolated.

[0091] As used herein, the term “DNA” refers to deoxyribonucleic acid. In various embodiments, the term DNA refers to genomic DNA, recombinant DNA, synthetic DNA, or cDNA. In one embodiment, DNA refers to genomic DNA or cDNA. In particular embodiments, the DNA comprises a “target region.” DNA libraries contemplated herein include genomic DNA libraries and cDNA libraries constructed from RNA, *e.g.*, an RNA expression library. In various embodiments, the DNA libraries comprise one or more additional DNA sequences and/or tags.

[0092] The terms “target genetic locus” and “DNA target region” are used interchangeably herein and refer to a region of interest within a DNA sequence. In various embodiments, targeted genetic analyses are performed on the target genetic locus. In particular embodiments, the DNA target region is a region of a gene that is associated with a particular genetic state, genetic condition, genetic diseases; fetal testing; genetic mosaicism, paternity testing; predicting response to drug treatment; diagnosing or monitoring a medical condition; microbiome profiling; pathogen screening; or organ transplant monitoring. In further embodiments, the DNA target region is a DNA sequence that is associated with a particular human chromosome, such as a particular autosomal or X-linked chromosome, or region thereof (*e.g.*, a unique chromosome region).

[0093] As used herein, the terms “circulating DNA,” “circulating cell-free DNA,” and “cell-free DNA” are often used interchangeably and refer to DNA that is extracellular DNA, DNA that has been extruded from cells, or DNA that has been released from necrotic or apoptotic cells. This term is often used in contrast to “cellular genomic DNA” or “cellular DNA,” which are used interchangeably herein and refer to genomic DNA that is contained within the cell (i.e. the nucleus) and is only accessible to molecular biological techniques such as those described herein, by lysing or otherwise disrupting the integrity of the cell.

[0094] A “subject,” “individual,” or “patient” as used herein, includes any animal that exhibits a symptom of a condition that can be detected or identified with compositions contemplated herein. Suitable subjects include laboratory animals (such as mouse, rat, rabbit, or guinea pig), farm animals (such as horses, cows, sheep, pigs), and domestic animals or pets (such as a cat or dog). In particular embodiments, the subject is a mammal. In certain embodiments, the subject is a non-human primate and, in preferred embodiments, the subject is a human.

[0095] As used herein, the term “paired” when used with respect to two different polynucleotide sequences or regions of DNA comprising different polynucleotide sequences, means that the two different polynucleotide sequences or regions of DNA comprising different polynucleotide sequences are present on the same polynucleotide. For example, if a particular sample tag region of DNA is said to be paired to particular amplification region of DNA, it is meant that the sample tag region and the amplification tag are present on the same DNA polynucleotide molecule.

C. METHODS OF COPY NUMBER ANALYSIS

[0096] In various embodiments, a method for copy number analysis of a DNA target region DNA is provided. In certain embodiments, copy number analysis is performed by generating a genomic DNA library of DNA library fragments that each contain genomic DNA fragment and an adaptor, isolating the DNA library fragments containing the DNA target regions, and performing a quantitative genetic analysis of the DNA target region. By “quantitative genetic analysis” it is meant an analysis performed by any molecular biological technique that is able to quantify changes in a DNA (*e.g.*, a gene, genetic locus, target region of interest, etc.) including but not limited to DNA mutations, SNPs, translocations, deletions,

and copy number variations (CNVs). In certain embodiments, the quantitative genetic analysis is performed by sequencing, for example, next generation sequencing.

[0097] Next-generation DNA sequencing (NGS) is ideally suited for two diagnostic applications. The first is the determination of DNA sequence on a vast scale. In the present context, this capability enables the search for rare, actionable variants that guide effective treatment decisions. The second is counting gene copy number. The output of millions of independent sequences can enable precise measurement of gene copy number on a genome-wide scale. The emergence of non-invasive prenatal testing for fetal trisomy from maternal blood samples is a testament to this capability. RNAseq, that is, the technology of gene expression profiling using NGS is another example, albeit the input is RNA (cDNA) rather than genomic DNA. Comparisons of current capture methods are described Samorodnitsky *et al.* J Mol Diagn. 2015 Jan;17(1):64-75.

[0098] The present invention extends NGS counting capability into the realm of targeted hybrid capture methods. The methods described here are effective for the detection of copy number variation at least in part because they possess the following four qualities:

(a) The present methods differentiate between unique clones and redundant clones. NGS sequencing of amplified genomic DNA library fragments results in a plurality of individual NGS reads, each comprising adaptor-encoded sequence information linked to a specific human genomic sequence. These elements define the identity of every clone. Because captured genomic regions are amplified by PCR, it is not uncommon for the same clone to be encountered several times in a subsequent NGS analysis. Groups of reads that are derived from a single cloning and capture process are termed “redundant reads.” Two or more redundant reads are identified as redundant reads based on the sequencing information provided by the unique molecular identification elements (UMIE). The UMIE refers to the combination of the sequence information from the adaptor tags and the start of the genomic DNA sequence. Two or more reads comprising identical UMIEs are identified as redundant reads. Redundant reads are grouped together and a single, representative consensus sequence is assembled from families of redundant reads. This consensus sequence is designated as a “unique read” or a “unique genomic sequence” (UGS). Each unique read represents a separate clone from the original DNA specimen. The process of identifying and grouping redundant clone families and of generating a single unique read representative of this family is defined as “deduplication.” The adaptors used to create genomic libraries possess a very

deep repertoire of unique sample tag information (15,360 codes per adaptor). When applied in conjunction with the exact mapping coordinates of each captured genomic clone (which can span >100 different positions relative to a capture probe), each unique clone that is generated in a genomic library and subsequently retrieved by a target-specific capture probe has an extremely high likelihood of being differentiable from all other unique clones that encompass the same capture environment. The ability to differentiate between unique clones and redundant clones is central to the methods described herein.

(b) The adaptors used to create genomic libraries permit sample multiplexing without creating adaptor-to-adaptor variability in copy number counts. A central foundation of copy number determination is the simultaneous analysis of a set of samples that have all been processed within a single sequencing run. This allows positive and negative controls to be included along with clinical samples. A major issue with previous adaptor design iterations induced subtle shifts in gene copy counts among identical control samples, in effect setting a signal-to-noise uncertainty threshold that was too high to be clinically useful in blood-based, solid tumor genotyping assays. The present invention overcomes this issue and substantially lowers the signal-to-noise threshold such that single copy gene loss is detectable at $\leq 2\%$ minor allele frequency. This improved signal recognition enables the methods of the present invention to have significant clinical utility in circulating tumor DNA assays.

(c) The proprietary targeted hybrid capture method used herein must produce highly uniform “on-target” read coverage across all targets. Methods that rely on counting of unique genomic fragments to estimate copy number, such as the ones described herein, must achieve near-saturation in terms of encountering all possible unique fragments. Near-saturation is only achieved by oversampling, that is to say, gathering more sequencing reads than the number of unique reads that will ultimately be encountered. To be practical, scalable, and economical, the unique reads in a targeted hybrid capture library must exhibit sufficient uniformity such that < 10-fold oversampling of on-target reads, and preferably < 4-fold oversampling of on-target reads will capture > 90% of unique on-target reads at all target loci.

(d) The targeted hybrid capture method (See U.S. Patent Publication No. 2014-0274731) must have high on-target capture rates. To be practical, scalable and economical, in other words to be a distinguishing feature of the present disclosure relative to other art in the field, the method must achieve >90%, preferably >95% on-target reads. With on-target

mapping rates exceeding 95%, the requirement for 4 to 10-fold oversampling of on-target reads and the requirement for overall oversampling are one in the same.

[0099] In some embodiments, the number of copies of the DNA target region present in the sample is determined by the quantitative genetic analysis. In some embodiments, the copy number of the DNA target region is determined by comparing the amount of copies of DNA target regions present in the sample and comparing it to amounts of DNA target regions present in one or more samples with known copy number.

[00100] Particular embodiments contemplate that the compositions and methods described herein are particularly useful for detecting changes in copy number in a sample of genomic DNA, where only a portion of the total genomic DNA in the sample has a change in copy number. For example, a significant tumor mutation may be present in a sample, *e.g.* a sample of cell free DNA, that is present in a minor allele frequency that is significantly less than 50% (*e.g.*, in the range of 0.1% to >20%), in contrast to conventional SNP genotyping where allele frequencies are generally ~100%, 50% or 0%. One of skill of the art will recognize that the compositions and methods of the current invention are also useful in detecting other types of mutation including single nucleotide variants (SNVs), short (*e.g.*, less than 40 base pairs (bp)) insertions, and deletions (indels), and genomic rearrangements including oncogenic gene fusions.

[00101] In certain embodiments, the compositions and/or methods of the present invention described herein are useful for, capable of, suited for, and/or able to detect, identify, observe, and/or reveal a change in copy number of one or more DNA target regions present in less than about 20%, less than about 19%, less than about 18%, less than about 17%, less than about 16%, less than about 15%, less than about 14%, less than about 13%, less than about 12%, less than about 11%, less than about 10%, less than about 9%, less than about 8%, less than about 7%, less than about 6%, less than about 5%, less than about 4%, less than about 3%, less than about 2%, less than about 1%, less than about 0.5%, less than about 0.2%, or less than about 0.1% of the total genomic DNA from the sample. In some embodiments, the methods of the present invention are useful for, capable of, suited for, and/or able to detect, identify, observe, and/or reveal a change in copy number of one or more DNA target regions present in between about 0.01% to about 100%, about 0.01% to about 50%, and or about 0.1% to about 20% of the total genomic DNA from the sample.

[00102] Particular embodiments are represented by the conceptual framework that is illustrated in FIG. 1. In FIG. 1, each gene is represented by a row and each patient

sample is represented as a column. Within any given genomic DNA sample, the number of fragments counted for each individual gene will have some variability, and that for any given DNA region of interest, *e.g.* a gene, perturbations in copy number are detected as significant fragment count deviations relative to the normalized counts to the DNA target region in other samples. Such an assay requires the gene-by-gene fragment counting profile within a sample to be reproducible, and also requires the sample-by-sample counting profiles to be highly comparable. Both assay requirements demand excellent signal-to-noise counting discrimination.

[00103] Some embodiments contemplate that the assay elements that contribute to increasing the signal to noise ratio are the genomic input, the number of probes, and the sequencing depth, as illustrated in FIG. 2.

[00104] In particular embodiments, a method for genetic analysis of cfDNA comprises: generating and amplifying a cfDNA library, determining the number of genome equivalents in the cfDNA library; and performing a quantitative genetic analysis of one or more genomic target loci.

[00105] Particular embodiments contemplate that the any of the methods and compositions described herein are effective for use to efficiently analyze, detect, diagnose, and/or monitor genetic states, genetic conditions, genetic diseases, genetic mosaicism, fetal diagnostics, paternity testing, microbiome profiling, pathogen screening, and organ transplant monitoring using genomic DNA, *e.g.*, cellular or cfDNA, where all or where only a portion of the total genomic DNA in the sample has a feature of interest, *e.g.* a genetic lesion, mutation, single nucleotide variant (SNV). In some embodiments, a feature of interest is a genetic feature associated with a disease or condition. For example, a significant tumor mutation may be present in a sample, *e.g.* a sample of cfDNA, that is present in a minor allele frequency that is significantly less than 50% (*e.g.* in the range of 0.1% to >20%), in contrast to conventional SNP genotyping where allele frequencies are generally ~100%, 50% or 0%.

[00106] In certain embodiments, the compositions and/or methods of the present invention described herein are useful for, capable of, suited for, and/or able to detect, identify, observe, and/or reveal a genetic lesion of one or more DNA target regions present in less than about 20%, less than about 19%, less than about 18%, less than about 17%, less than about 16%, less than about 15%, less than about 14%, less than about 13%, less than about 12%, less than about 11%, less than about 10%, less than about 9%, less than about 8%, less than about 7%, less than about 6%, less than about 5%, less than about 4%, less than about

3%, less than about 2%, less than about 1%, less than about 0.5%, less than about 0.2%, or less than about 0.1% of the total genomic DNA from the sample. In some embodiments, the methods of the present invention are useful for, capable of, suited for, and/or able to detect, identify, observe, and/or reveal a genetic lesion of one or more DNA target regions present in between about 0.01% to about 100%, about 0.01% to about 50%, and or about 0.1% to about 20% of the total genomic DNA from the sample.

1. GENERATING A DNA LIBRARY

[00107] In particular embodiments, methods of genetic analysis contemplated herein comprise generating a DNA library comprising treating cfDNA or fragmented cellular genomic DNA with one or more end-repair enzymes to generate end-repaired DNA and attaching one or more adaptors to each end of the end-repaired DNA to generate the DNA library. Genomic DNA

[00108] In particular embodiments, the methods and compositions contemplated herein are designed to efficiently analyze, detect, diagnose, and/or monitor change in copy number using genomic DNA as an analyte. In certain embodiments, copy number analysis is performed by generating a genomic DNA library from genomic DNA obtained from a test sample, *e.g.*, a biological sample such as a tissue biopsy. In certain embodiments, the genomic DNA is circulating or cell free DNA. In some embodiments, the genomic DNA is cellular genomic DNA.

[00109] In certain embodiments, genomic DNA is obtained from a tissue sample or biopsy taken from a tissue, including but not limited to, bone marrow, esophagus, stomach, duodenum, rectum, colon, ileum, pancreases, lung, liver, prostate, brain, nerves, meningeal tissue, renal tissue, endometrial tissue, cervical tissue, breast, lymph node, muscle, and skin. In certain embodiments, the tissue sample is a biopsy of a tumor or a suspected tumor. In particular embodiments, the tumor is cancerous or suspected of being cancerous. In particular embodiments, the tissue sample comprises cancer cells or cells suspected of being cancerous.

[00110] Methods for purifying genomic DNA from cells or from a biologic tissue comprised of cells are well known in the art, and the skilled artisan will recognize optimal procedures or commercial kits depending on the tissue and the conditions in which the tissue is obtained. Some embodiments contemplate that purifying cellular DNA from a

tissue will require cell disruption or cell lysis to expose the cellular DNA within, for example by chemical and physical methods such as blending, grinding or sonicating the tissue sample; removing membrane lipids by adding a detergent or surfactants which also serves in cell lysis, optionally removing proteins, for example by adding a protease; removing RNA, for example by adding an RNase; and DNA purification, for example from detergents, proteins, salts and reagents used during cell lysis step. DNA purification may be performed by precipitation, for example with ethanol or isopropanol; by phenol–chloroform extraction.

[00111] In particular embodiments, cellular DNA obtained from tissues and/or cells are fragmented prior to and or during obtaining, generating, making, forming, and/or producing a genomic DNA library as described herein. One of skill in the art will understand that there are several suitable techniques for DNA fragmentation, and is able to recognize and identify suitable techniques for fragmenting cellular DNA for the purposes of generating a genomic DNA library for DNA sequencing, including but not limited to next-generation sequencing. Certain embodiments contemplate that cellular DNA can be fragmented into fragments of appropriate and/or sufficient length for generating a library by methods including but not limited to physical fragmentation, enzymatic fragmentation, and chemical shearing.

[00112] Physical fragmentation can include, but is not limited to, acoustic shearing, sonication, and hydrodynamic shear. In some embodiments, cellular DNA is fragmented by physical fragmentation. In particular embodiments, cellular DNA is fragmented by acoustic shearing or sonication. Particular embodiments contemplate that acoustic shearing and sonication are common physical methods used to shear cellular DNA. The Covaris® instrument (Woburn, MA) is an acoustic device for breaking DNA into 100-5kb bp. Covaris also manufactures tubes (gTubes) which will process samples in the 6-20 kb for Mate-Pair libraries. The Bioruptor® (Denville, NJ) is a sonication device utilized for shearing chromatin, DNA and disrupting tissues. Small volumes of DNA can be sheared to 150-1kb in length. Hydroshear from Digilab (Marlborough, MA) utilizes hydrodynamic forces to shear DNA. Nebulizers (Life Tech, Grand Island, NY) can also be used to atomize liquid using compressed air, shearing DNA into 100-3kb fragments in seconds. Nebulization is low cost, but the process can cause a loss of about 30% of the cellular DNA from the original sample. In certain embodiments, cellular DNA is fragmented by sonication.

[00113] Enzymatic fragmentation can include, but is not limited to, treatment with a restriction endonuclease, *e.g.* DNase I, or treatment with a nonspecific nuclease. In

some embodiments, cellular DNA is fragmented by enzymatic fragmentation. In particular embodiments, the cellular DNA is fragmented by treatment with a restriction endonuclease. In some embodiments, the cellular DNA is fragmented by treatment with a nonspecific nuclease. In certain embodiments, the cellular DNA is fragmented by treatment with a transposase. Certain embodiments contemplate that enzymatic methods to shear cellular DNA into small pieces include DNase I, a combination of maltose binding protein (MBP)-T7 Endo I and a non-specific nuclease *Vibrio vulnificus* (Vvn) New England Biolabs's (Ipswich, MA) Fragmentase and Nextera tagmentation technology (Illumina, San Diego, CA). The combination of non-specific nuclease and T7 Endo synergistically work to produce non-specific nicks and counter nicks, generating fragments that disassociate 8 nucleotides or less from the nick site. Tagmentation uses a transposase to simultaneously fragment and insert adapters onto double stranded DNA.

[00114] Chemical fragmentation can include treatment with heat and divalent metal cation. In some embodiments, genomic DNA is fragmented by chemical fragmentation. Particular embodiments contemplate that chemical shear is more commonly used for the breakup of long RNA fragments as opposed to genomic DNA. Chemical fragmentation is typically performed through the heat digestion of DNA with a divalent metal cation (magnesium or zinc). The length of DNA fragments can be adjusted by increasing or decreasing the time of incubation.

[00115] In particular embodiments, the methods and compositions contemplated herein are designed to efficiently analyze, detect, diagnose, and/or monitor change in copy number using cell-free DNA (cfDNA) as an analyte. The size distribution of cfDNA ranges from about 150 bp to about 180 bp fragments. Fragmentation of cfDNA may be the result of endonucleolytic and/or exonucleolytic activity and presents a formidable challenge to the accurate, reliable, and robust analysis of cfDNA. Another challenge for analyzing cfDNA is its short half-life in the blood stream, on the order of about 15 minutes. Without wishing to be bound to any particular theory, the present invention contemplates, in part, that analysis of cfDNA is like a "liquid biopsy" and is a real-time snapshot of current biological processes.

[00116] Moreover, because cfDNA is not found within cells and may be obtained from a number of suitable sources including, but not limited to, biological fluids and stool samples, it is not subject to the existing limitations that plague next generation sequencing analysis, such as direct access to the tissues being analyzed.

[00117] Illustrative examples of biological fluids that are suitable sources from which to isolate cfDNA in particular embodiments include, but are not limited to amniotic fluid, blood, plasma, serum, semen, lymphatic fluid, cerebral spinal fluid, ocular fluid, urine, saliva, mucous, and sweat. In particular embodiments, the biological fluid is blood or blood plasma.

[00118] In certain embodiments, commercially available kits and other methods known to the skilled artisan can be used to isolate cfDNA directly from the biological fluids of a subject or from a previously obtained and optionally stabilized biological sample, *e.g.*, by freezing and/or addition of enzyme chelating agents including, but not limited to EDTA, EGTA, or other chelating agents specific for divalent cations.

(a) Generating End-Repaired cfDNA

[00119] In particular embodiments, generating a genomic DNA library comprises the end-repair of isolated cfDNA or fragmented cellular DNA. The fragmented cfDNA or cellular DNA is processed by end-repair enzymes to generate end-repaired cfDNA with blunt ends, 5'-overhangs, or 3'-overhangs. In some embodiments, the end-repair enzymes can yield for example. In some embodiments, the end-repaired cfDNA or cellular DNA contains blunt ends. In some embodiments, the end-repaired cellular DNA or cfDNA is processed to contain blunt ends. In some embodiments, the blunt ends of the end-repaired cfDNA or cellular DNA are further modified to contain a single base pair overhang. In some embodiments, end-repaired cfDNA or cellular DNA containing blunt ends can be further processed to contain adenine (A)/thymine (T) overhang. In some embodiments, end-repaired cfDNA or cellular DNA containing blunt ends can be further processed to contain adenine (A)/thymine (T) overhang as the single base pair overhang. In some embodiments, the end-repaired cfDNA or cellular DNA has non-templated 3' overhangs. In some embodiments, the end-repaired cfDNA or cellular DNA is processed to contain 3' overhangs. In some embodiments, the end-repaired cfDNA or cellular DNA is processed with terminal transferase (TdT) to contain 3' overhangs. In some embodiments, a G-tail can be added by TdT. In some embodiments, the end-repaired cfDNA or cellular DNA is processed to contain overhang ends using partial digestion with any known restriction enzymes (*e.g.*, with the enzyme Sau3A, and the like).

(b) Attaching Adaptor Molecules to End-Repaired cfDNA

[00120] In particular embodiments, generating a cfDNA library comprises attaching one or more adaptors to each end of the end-repaired cfDNA. The present invention contemplates, in part, an adaptor module designed to accommodate large numbers of genome equivalents in cfDNA libraries. Adaptor modules are configured to measure the number of genome equivalents present in cfDNA libraries, and, by extension, the sensitivity of sequencing assays used to identify sequence mutations.

[00121] As used herein, the terms “adaptor” and “adaptor module” are used for interchangeably, and refer to a polynucleotide comprising that comprises at least three elements: an amplification region, a sample tag region, and an anchor region. In particular embodiments, the adaptor comprises an amplification region, a sample tag region, and an anchor region. In some embodiments, the adaptor also comprises a unique molecule identifier (UMI). In particular embodiments, the adaptor comprises one or more amplification regions, one or more sample tag regions, one or more UMIs, and/or one or more anchor regions. In some embodiments, the adaptor comprises, in order from 5' to 3', an amplification region, a sample tag region, a UMI, and an anchor region. In particular embodiments, the adaptor comprises, in order from 5' to 3', an amplification region, a sample tag region, a UMI, and an anchor region. In certain embodiments, the UMI is contained within the sample tag region, and the adaptor comprises, in order from 5' to 3', an amplification region, an integrated sample tag/UMI region, and an anchor region.

[00122] As used herein, the term “amplification region” refers to an element of the adaptor molecule that comprises a polynucleotide sequence capable of serving as a primer recognition site for PCR amplification. In particular embodiments, an adaptor comprises an amplification region that comprises one or more primer recognition sequences for single-primer amplification of a genomic DNA library. In some embodiments, the amplification region comprises one, two, three, four, five, six, seven, eight, nine, ten, or more primer recognition sequences for single-primer amplification of a genomic DNA library.

[00123] In some embodiments, the amplification region is about is between 5 and 50 nucleotides, between 10 and 45 nucleotides, between 15 and 40 nucleotides, or between 20 and 30 nucleotides in length. In some embodiments, the amplification region is 10 nucleotides, 11 nucleotides, 12 nucleotides, 13 nucleotides, 14 nucleotides, 15 nucleotides, 16 nucleotides, 17 nucleotides, about 18 nucleotides, 19 nucleotides, 20 nucleotides, 21 nucleotides, 22 nucleotides, 23 nucleotides, 24 nucleotides, 25 nucleotides, 26 nucleotides, 27

nucleotides, 28 nucleotides, 29 nucleotides, 30 nucleotides, 31 nucleotides, 32 nucleotides, 33 nucleotides, 34 nucleotides, 35 nucleotides, 36 nucleotides, 37 nucleotides, 38 nucleotides, 39 nucleotides, or 40 nucleotides or more. In particular embodiments, the amplification region is 25 nucleotides in length.

[00124] As used herein, the term “sample tag” or sample tag region” are used interchangeably and refer to an element of the adaptor that comprises a polynucleotide sequence that uniquely identifies the particular DNA fragment as well as the sample from which it was derived.

[00125] In certain embodiments, the sample tag region is about is between 3 and 50 nucleotides, between 3 and 25 nucleotides, or between 5 and 15 nucleotides in length. In some embodiments, the sample tag region is 3 nucleotides, 4 nucleotides, 5 nucleotides, 6 nucleotides, 7 nucleotides, 8 nucleotides, 9 nucleotides, 10 nucleotides, about 11 nucleotides, 12 nucleotides, 13 nucleotides, 14 nucleotides, 15 nucleotides, 16 nucleotides, 17 nucleotides, 18 nucleotides, 19 nucleotides, or 20 nucleotides or more in length.

[00126] In certain embodiments, the adaptor comprises a UMI multiplier, wherein the UMI multiplier is 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, or at least 10 nucleotides in length.

[00127] In certain embodiments, each nucleotide position of the UMI multiplier can comprise any of adenine, guanine, cytosine, or thymine. Thus, in some embodiments, a UMI multiplier comprising n number of nucleotides can comprise any of n^4 possible nucleotide sequences. In some embodiments, the UMI multiplier is one nucleotide in length and comprises one of four possible sequences. In some embodiments, the UMI multiplier is two nucleotides in length and comprises one of sixteen possible sequences. In some embodiments, the UMI multiplier is three nucleotides in length and comprises one of 64 possible sequences. In some embodiments, the UMI multiplier is four nucleotides in length and comprises one of 256 possible sequences. In some embodiments, the UMI multiplier is five nucleotides in length and comprises one of 1,024 possible sequences. In some embodiments, the UMI multiplier is six nucleotides in length and comprises one of 4,096 possible sequences. In some embodiments, the UMI multiplier is seven nucleotides in length and comprises one of 16,384 possible sequences. In some embodiments, the UMI multiplier is eight nucleotides in length and comprises one of 65,536 possible sequences. In some embodiments, the UMI multiplier is nine nucleotides in length and comprises one of 262,144

possible sequences. In some embodiments, the UMI multiplier is ten or more nucleotides in length and comprises one of 1,048,576 or more possible sequences.

[00128] In particular embodiments, the adaptor comprises a UMI multiplier, wherein the UMI multiplier is adjacent to or contained within the sample tag region (FIG. 5A). Illustrative examples of UMI multipliers adjacent or contained within the sample tag are shown in FIG. 5B. In FIG. 5B, an 8-mer sample tag region is shown with an adjacent UMI multiplier (top and bottom rows) or a UMI multiplier incorporated within the sample tag (middle 7 rows). In some embodiments, that adaptor comprises a sample tag that is eight nucleotides in length and a UMI multiplier that is three nucleotides in length and comprises one of 64 possible sequences, and wherein the UMI multiplier is adjacent to or contained within the sample tag region. In some embodiments, identical processes attach full length adaptor to the other end of the genomic fragments.

[00129] In particular embodiments, an adaptor module comprises one or more anchor sequences. As used herein, an “anchor region” and “anchor sequence” are used interchangeably and refer to a nucleotide sequence that hybridizes to a partner oligonucleotide. In some embodiments, the anchor region comprises the following three properties: (1) each anchor sequence is part of a family of two or more anchor sequences that collectively represent each of the four possible DNA bases at each site within extension; this feature, balanced base representation, is useful to calibrate proper base calling in sequencing reads in particular embodiments; (2) each anchor sequence is composed of only two of four possible bases, and these are specifically chosen to be either an equal number of A + C or an equal number of G + T; an anchor sequence formed from only two bases reduces the possibility that the anchor sequence will participate in secondary structure formation that would preclude proper adaptor function; and (3) because each anchor sequence is composed of equal numbers of A + C or G + T, each anchor sequence shares roughly the same melting temperature and duplex stability as every other anchor sequence in a set of four.

[00130] In some embodiments, the anchor sequences is between 1 and 50 nucleotides in length. In some embodiments, the anchor sequences is between 4 and 40 nucleotides in length. In certain embodiments, the anchor region is between 5 and 25 nucleotides in length. In particular embodiments, the anchor region is at least 4 nucleotides, at least six nucleotides, at least 8 nucleotides, at least 10 nucleotides, at least 12 nucleotides, at least 14 nucleotides, or at least 16 nucleotides in length. In particular embodiments, the anchor region is 10 nucleotides in length.

[00131] In particular embodiments, an attachment step comprises attaching/ligating an adaptor module to the end-repaired cfDNA or cellular DNA to generate a “tagged” genomic DNA library. In some embodiments, a single adaptor module is employed. In some embodiments, two, three, four or five adaptor modules are employed. In some embodiments, an adaptor module of identical sequence is attached to each end of the fragmented end-repaired DNA.

[00132] In some embodiments, a plurality of adaptor species is attached to an end-repaired cellular or cell free genomic DNA fragments. Each of the plurality of adaptors may comprise one or more amplification regions for the amplification of the cfDNA or cellular DNA library, one or more sample tag regions for the identification of the cfDNA or cellular genomic DNA fragment and identification of the individual sample; and one or more sequences for DNA sequencing.

[00133] In some embodiments, a plurality of adaptor species is attached to an end-repaired cellular or cell free genomic DNA fragments of a sample, and the plurality of adaptors all comprise amplification regions of an identical nucleotide sequence.

[00134] In certain embodiments, the genomic DNA from a sample is attached with a plurality of adaptors that comprise sample tag sequences that all are different from other sequences of sample tag regions in adaptors that are attached to genomic DNA fragments from other samples.

[00135] In particular embodiments, a plurality of adaptor species is attached to an end-repaired cellular or cell free genomic DNA fragments from a sample, and the plurality of adaptors all comprise one or more sample tag regions comprising one of between 2 and 10,000 nucleotide sequences, one of between 5 and 5,000 nucleotide sequences, one of between 25 and 1,000 nucleotide sequences, one of between 50 and 500 nucleotide sequences, one of between 100 and 400 nucleotide sequences, or one of between 200 and 300 nucleotide sequences. In some embodiments, the sample tag region of each adaptor is 8 nucleotides in length, and each sample tag region of the plurality of adaptors comprises one of 240 nucleotide sequences.

[00136] In certain embodiments, a plurality of adaptor species is attached to an end-repaired cellular or cell free genomic DNA fragments from a sample, and the sample tag regions of the plurality of adaptors comprises nucleotide sequences that are different from each other by a Hamming distance of 1, 2, 3, 4 or greater than 4. In particular embodiments, the Hamming distance is 2.

[00137] In particular embodiments, the sample tag regions of the plurality of adaptors that are attached to genomic DNA fragments of a sample are 8 nucleotides in length, and comprise one of 240 nucleotide sequences that are different from each other by a Hamming distance of 2.

[00138] In certain embodiments, the sample tag region serves to identify individual genomic DNA fragments and to identify the individual sample, i.e., the genomic library source. For example, when the sample tags of a plurality of adaptors attached to a sample have one of 240 possible sequences, each sample is identified as having one of 240 possible tags, and each sample receives a set of 240 tags that are discrete from any other sample by Hamming distance of two (meaning two base changes are required to change one tag into another). These same tags are used to enumerate clone diversity and thus they also serve as sequence tags, i.e., to identify genomic DNA fragments. To further augment the diversity of possible sequence tags, UMI multipliers may be added. For example, a UMI multiplier can be added to the adaptor region comprising 3 nucleotides consisting of the 64 possible combinations of 3 bases. In addition, the plurality of adaptors can comprise more than one anchor sequence. For example, a plurality of adaptors may contain 4 different anchor sequences are used simultaneously. These anchor sequences may also be used during sample de-multiplexing to lower errors.

[00139] FIG. 4 shows an illustrative comparison between a first generation adaptor (FIG. 4A and 4B) and an adaptor of the present invention (FIG. 4C – FIG. 4E). FIG. 4A and FIG. 4B show an example of first generation adaptor that is 40 nt in length and consisted of a discrete PCR amplification sequence, sequence tag, and sample tag. Here, the sample is identified by a fixed sequence (sequence tag) that is present on all adaptors that are used to generate a DNA library from the sample. Individual genomic fragments are identified by a separate and distinct sequences (sequence tag). FIG. 4C – FIG. 4E show an illustrative example of an adaptor from the present invention. The illustrative adaptor shown is 47 nucleotides in length, and the sequence tag is combined with the sample tag. There is an additional 3 nt sequence, the UMI multiplier, consisting of the 64 possible combinations of 3 bases. The 10 nt anchor sequence is one of four different distinct sequences.

[00140] Thus, in the illustrative example (*See* FIG. 4C – FIG. 4E), a set of adaptors that are used in connection with a single sample comprise 240 sample tag sequences that can be split into four sets of sample tag sequences with each set comprising 60 tags (one for each nucleotide, A, C, T and G). Thus, each set of 60 tags is specific to one of four anchor

sequences. In total, a pool of 240 possible sample tag configurations are possible per sample. Specifically, in this scenario, the 240 sample tag sequences are divided into four sets of 60 sequences, with each set directed to a specific anchor region. Therefore, the sample ID involves not only the sequence information from the eight nucleotide sample tag, but also the associated anchor sequence information. In addition, the position of sequences within the read is fixed, and therefore the sample tags and anchor sequences must have a fixed position within a sequencing read in order to pass inclusion filters for downstream consideration. Further, the inclusion of the UMI multiplier increases the sequence tag diversity from 240 to $240 \times 64 = 15,360$ possible sequence tags.

[00141] Attachment of one or more adaptors contemplated herein may be carried out by methods known to those of ordinary skill in the art. In particular embodiments, one or more adaptors contemplated herein are attached to end-repaired cfDNA that comprises blunt ends. In certain embodiments, one or more adaptors contemplated herein are attached to end-repaired cfDNA that comprises complementary ends appropriate for the attachment method employed. In certain embodiments, one or more adaptors contemplated herein are attached to end-repaired cfDNA that comprises a 3' overhang.

[00142] In some embodiments, attaching the genomic DNA fragments to a plurality of adaptors includes the steps of attaching the end repaired cfDNA or cellular DNA fragments to an oligonucleotide containing at least a portion of an anchor region. In some embodiments, the oligonucleotide contains the whole anchor region. In particular embodiments, the oligonucleotide is a DNA duplex comprising a 5' phosphorylated attachment strand duplexed with a partner strand, wherein the partner strand is blocked from attachment by chemical modification at its 3' end, and wherein the attachment strand is attached to the genomic DNA fragment. In certain embodiments, the DNA fragments attached with at least a portion of the anchor region are then annealed with DNA oligonucleotides encoding the full length adaptor sequences. In particular embodiments, one or more polynucleotide kinases, one or more DNA ligases, and/or one or more DNA polymerases are added to the genomic DNA fragments and the DNA oligonucleotides encoding the full length adaptor sequence. In some embodiments, the polynucleotide kinase is T4 polynucleotide kinase. In some embodiments, the DNA ligase is Taq DNA ligase. In certain embodiments, the DNA polymerase is Taq polymerase. In particular embodiments, the DNA polymerase is full length Bst polymerase.

[00143] FIG. 6 shows an illustrative method for attaching a plurality of adaptors to the 3' end of repaired DNA fragments. In the first step, the anchor sequence is attached to the 3' ends of genomic fragments. In this step, the anchor portion is a DNA duplex in which the ten nucleotide 5' phosphorylated "attachment strand" is duplexed with an eight nucleotide "partner strand" that is blocked from attachment by chemical modification at its 3' end. The anchor duplex is blunt-ended on the phosphorylated/blocked end and can therefore attach to blunt-ended genomic fragments. In the next step, pools of oligonucleotides encoding the full adaptor sequences are annealed to the initial anchor sequence. The combined action of T4 polynucleotide kinase, Taq DNA ligase, and full-length Bst polymerase attach this oligonucleotide via ligation as illustrated for the top strand and extend the initial anchor sequence by DNA polymerization on the bottom strand to complete the full-length adaptor sequence. Identical processes may be used to attach full length adaptors to the 5' end of the genomic fragments.

2. DNA LIBRARY AMPLIFICATION

[00144] In particular embodiments, methods of genetic analysis contemplated herein comprise amplification of a genomic DNA library, *e.g.* a cellular DNA library or a cfDNA library, to generate a DNA clone library or a library of DNA clones, *e.g.*, a cfDNA clone library or a library of cfDNA clones, or a cellular DNA clone library or a library of cellular DNA clones. Each molecule of the DNA library comprises an adaptor attached to each end of an end-repaired DNA fragments, and each adaptor comprises one or more amplification regions. In some embodiments, different adaptors are attached to different ends of the end-repaired cfDNA. In particular embodiments, different adaptors are attached to different ends of the end-repaired cellular DNA.

[00145] In some embodiments, the same adaptor is attached to both ends of the DNA fragment. Attachment of the same adaptor to both ends of end-repaired DNA allows for PCR amplification with a single primer sequence. In particular embodiments, a portion of the adaptor attached-cfDNA library will be amplified using standard PCR techniques with a single primer sequence driving amplification. In one embodiment, the single primer sequence is about 25 nucleotides, optionally with a projected T_m of $\geq 55^\circ \text{C}$ under standard ionic strength conditions.

[00146] In particular embodiments, picograms of the initial genomic DNA library, *e.g.* a cellular DNA library or cfDNA library, are amplified into micrograms of DNA clones, implying a 10,000-fold amplification. The amount of amplified product can be measured using methods known in the art, *e.g.*, quantification on a Qubit 2.0 or Nanodrop instrument.

3. DETERMINING THE NUMBER OF GENOME EQUIVALENTS

[00147] In various embodiments, a method for genetic analysis of genomic DNA comprises determining the number of genome equivalents in the DNA clone library. As used herein, the term “genome equivalent” refers to the number of genome copies in each library. An important challenge met by the compositions and methods contemplated herein is achieving sufficient assay sensitivity to detect and analysis rare genetic mutations or differences in genetic sequence. To determine assay sensitivity value on a sample-by-sample basis, the numbers of different and distinct sequences that are present in each sample are measured by measuring the number of genome equivalents that are present in a sequencing library. To establish sensitivity, the number of genome equivalents must be measured for each sample library.

[00148] The number of genome equivalents can be determined by qPCR assay or by using bioinformatics-based counting after sequencing is performed. In the process flow of clinical samples, qPCR measurement of genome equivalents is used as a QC step for DNA libraries, *e.g.*, cfDNA libraries or genomic DNA libraries. It establishes an expectation for assay sensitivity prior to sequence analysis and allows a sample to be excluded from analysis if its corresponding DNA clone library lacks the required depth of genome equivalents. Ultimately, the bioinformatics-based counting of genome equivalents is also used to identify the genome equivalents – and hence the assay sensitivity and false negative estimates – for each given DNA clone library.

[00149] The empirical qPCR assay and statistical counting assays should be well correlated. In cases where sequencing fails to reveal the sequence depth in a DNA clone library, reprocessing of the DNA clone library and/or additional sequencing may be required.

[00150] In one embodiment, the genome equivalents in a cellular DNA or cfDNA clone library are determined using a quantitative PCR (qPCR) assay. In a particular embodiment, a standard library of known concentration is used to construct a standard curve

and the measurements from the qPCR assay are fit to the resulting standard curve and a value for genome equivalents is derived from the fit. The present inventors have discovered that a qPCR “repeat-based” assay comprising one primer that specifically hybridizes to a common sequence in the genome, *e.g.*, a repeat sequence, and another primer that binds to the primer binding site in the adaptor, measured an 8-fold increase in genome equivalents compared to methods using just the adaptor specific primer (present on both ends of the cfDNA clone). The number of genome equivalents measured by the repeat-based assays provides a more consistent library-to-library performance and a better alignment between qPCR estimates of genome equivalents and bioinformatically counted tag equivalents in sequencing runs.

[00151] Illustrative examples of repeats suitable for use in the repeat-based genome equivalent assays contemplated herein include, but not limited to: short interspersed nuclear elements (SINEs), *e.g.*, Alu repeats; long interspersed nuclear elements (LINEs), *e.g.*, LINE1, LINE2, LINE3; microsatellite repeat elements, *e.g.*, short tandem repeats (STRs), simple sequence repeats (SSRs); and mammalian-wide interspersed repeats (MIRs).

[00152] In one embodiment, the repeat is an Alu repeat.

4. QUANTITATIVE GENETIC ANALYSIS

[00153] In various embodiments, a method for genetic analysis of genomic DNA, *e.g.*, genomic cellular or cfDNA, comprises quantitative genetic analysis of one or more target genetic loci of the DNA library clones. Quantitative genetic analysis comprises one or more of, or all of, the following steps: capturing DNA clones comprising a target genetic locus; amplification of the captured targeted genetic locus; sequencing of the amplified captured targeted genetic locus; and bioinformatic analysis of the resulting sequence reads. As used herein, the terms “DNA library clone” refer to a DNA library fragment wherein the combination of the adaptor and the genomic DNA fragment result in a unique DNA sequence (*e.g.*, a DNA sequence that can be distinguished from that of another DNA library clone).

(a) Capture of Target Genetic Locus

[00154] The present invention contemplates, in part, a capture probe module designed to retain the efficiency and reliability of larger probes but that minimizes

uninformative sequence generation in a genomic DNA library that comprises smaller DNA fragments, *e.g.*, a cfDNA clone library. A “capture probe” or “capture probe module” as used herein, are used interchangeably and refer to a polynucleotide that comprises a capture probe sequence and a tail sequence. In particular embodiments, the capture probe module sequence or a portion thereof serves as a primer binding site for one or more sequencing primers.

[00155] In particular embodiments, a capture probe module comprises a capture probe. As used herein a “capture probe” refers to a region capable of hybridizing to a specific DNA target region. In some embodiments, the capture probes are used with genomic DNA library constructed from cellular DNA. In particular embodiments, the capture probes are used with genomic DNA library constructed from cfDNA. Because the average size of cfDNA is about 150 to about 170 bp and is highly fragmented, certain embodiments are directed compositions and methods contemplated herein comprise the use of high density and relatively short capture probes to interrogate DNA target regions of interest. In some embodiments, the capture probes are capable of hybridizing to DNA target regions that are distributed across all chromosomal segments at a uniform density. A set of such capture probes is referred to herein as “chromosomal stability probes.” Chromosomal stability probes are used to interrogate copy number variations on a genome-wide scale in order to provide a genome-wide measurement of chromosomal copy number (*e.g.*, chromosomal ploidy).

[00156] One particular concern with using high density capture probes is that generally capture probes are designed using specific “sequence rules.” For example, regions of redundant sequence or that exhibit extreme base composition biases are generally excluded in designing capture probes. However, the present inventors have discovered that the lack of flexibility in capture probe design rules does not substantially impact probe performance. In contrast, capture probes chosen strictly by positional constraint provided on-target sequence information; exhibit very little off-target and unmappable read capture; and yield uniform, useful, on-target reads with only few exceptions. Moreover, the high redundancy at close probe spacing more than compensates for occasional poor-performing capture probes.

[00157] In particular embodiments, a target region is targeted by a plurality of capture probes, wherein any two or more capture probes are designed to bind to the target region within 10 nucleotides of each other, within 15 nucleotides of each other, within 20 nucleotides of each other, within 25 nucleotides of each other, within 30 nucleotides of each other, within 35 nucleotides of each other, within 40 nucleotides of each other, within 45

nucleotides of each other, or within 50 nucleotides or more of each other, as well as all intervening nucleotide lengths.

[00158] In one embodiment, the capture probe is about 25 nucleotides, about 26 nucleotides, about 27 nucleotides, about 28 nucleotides, about 29 nucleotides, about 30 nucleotides, about 31 nucleotides, about 32 nucleotides, about 33 nucleotides, about 34 nucleotides, about 35 nucleotides, about 36 nucleotides, about 37 nucleotides, about 38 nucleotides, about 39 nucleotides, about 40 nucleotides, about 41 nucleotides, about 42 nucleotides, about 43 nucleotides, about 44 nucleotides, or about 45 nucleotides.

[00159] In one embodiment, the capture probe is about 100 nucleotides, about 200 nucleotides, about 300 nucleotides, about 400 nucleotides, or about 100 nucleotides. In another embodiment, the capture probe is from about 100 nucleotides to about 500 nucleotides, about 200 nucleotides to about 500 nucleotides, about 300 nucleotides to about 500 nucleotides, or about 400 nucleotides to about 500 nucleotides, or any intervening range thereof.

[00160] In a particular embodiment, the capture probe is 60 nucleotides. In another embodiment, the capture probe is substantially smaller than 60 nucleotides but hybridizes comparably, as well as, or better than a 60 nucleotide capture probe targeting the same DNA target region. In a certain embodiment, the capture probe is 40 nucleotides.

[00161] In certain embodiments, a capture probe module comprises a tail sequence. As used herein, the term “tail sequence” refers to a polynucleotide at the 5' end of the capture probe module, which in particular embodiments can serve as a primer binding site. In particular embodiments, a sequencing primer binds to the primer binding site in the tail region.

[00162] In particular embodiments, the tail sequence is about 5 to about 100 nucleotides, about 10 to about 100 nucleotides, about 5 to about 75 nucleotides, about 5 to about 50 nucleotides, about 5 to about 25 nucleotides, or about 5 to about 20 nucleotides. In certain embodiments, the third region is from about 10 to about 50 nucleotides, about 15 to about 40 nucleotides, about 20 to about 30 nucleotides or about 20 nucleotides, or any intervening number of nucleotides.

[00163] In particular embodiments, the tail sequence is about 30 nucleotides, about 31 nucleotides, about 32 nucleotides, about 33 nucleotides, about 34 nucleotides, about 35 nucleotides, about 36 nucleotides, about 37 nucleotides, about 38 nucleotides, about 39 nucleotides, or about 40 nucleotides.

[00164] In various embodiments, the capture probe module comprises a specific member of a binding pair to enable isolation and/or purification of one or more captured fragments of a tagged and or amplified genomic DNA library (*e.g.*, a cellular or cfDNA library) that hybridizes to the capture probe. In particular embodiments, the capture probe module is conjugate to biotin or another suitable hapten, *e.g.*, dinitrophenol, digoxigenin.

[00165] In various embodiments, the capture probe module is hybridized to a tagged and optionally amplified DNA library to form a complex. In some embodiments, the multifunctional capture probe module substantially hybridizes to a specific genomic target region in the DNA library.

[00166] Hybridization or hybridizing conditions can include any reaction conditions where two nucleotide sequences form a stable complex; for example, the tagged DNA library and capture probe module forming a stable tagged DNA library—capture probe module complex. Such reaction conditions are well known in the art and those of skill in the art will appreciate that such conditions can be modified as appropriate, *e.g.*, decreased annealing temperatures with shorter length capture probes, and within the scope of the present invention. Substantial hybridization can occur when the second region of the capture probe complex exhibits 100%, 99%, 98%, 97%, 96%, 95%, 94%, 93%, 92%, 91%, 90%, 89%, 88%, 85%, 80%, 75%, or 70% sequence identity, homology or complementarity to a region of the tagged DNA library.

[00167] In particular embodiments, the capture probe is about 40 nucleotides and has an optimal annealing temperature of about 44° C to about 47° C.

[00168] In certain embodiments, the methods contemplated herein comprise isolating a tagged cfDNA library—capture probe module complex. In particular embodiments, methods for isolating DNA complexes are well known to those skilled in the art and any methods deemed appropriate by one of skill in the art can be employed with the methods of the present invention (Ausubel *et al.*, *Current Protocols in Molecular Biology*, 2007-2012). In particular embodiments, the complexes are isolated using biotin—streptavidin isolation techniques.

[00169] In particular embodiments, removal of the single stranded 3'-ends from the isolated tagged DNA library fragments-capture probe module complex is contemplated. In certain embodiments, the methods comprise 3'-5' exonuclease enzymatic processing of the

isolated tagged DNA library-multifunctional capture probe module complex to remove the single stranded 3' ends.

[00170] In certain other embodiments, the methods comprise performing 5'-3' DNA polymerase extension of multifunctional capture probe utilizing the isolated tagged DNA library fragments as template.

[00171] In certain other embodiments, the methods comprise creating a hybrid capture probe-isolated tagged DNA target molecule, *e.g.*, a tagged cfDNA target molecule or a tagged cellular DNA target molecule, through the concerted action of a 5' FLAP endonuclease, DNA polymerization and nick closure by a DNA ligase.

[00172] A variety of enzymes can be employed for the 3'-5' exonuclease enzymatic processing of the isolated tagged DNA library-multifunctional capture probe module complex. Illustrative examples of suitable enzymes, which exhibit 3'-5' exonuclease enzymatic activity, that can be employed in particular embodiments include, but are not limited to: T4 or Exonucleases I, III, V (*See also*, Shevelev IV, Hübscher U., *Nat Rev Mol Cell Biol.* 3(5):364-76 (2002)). In particular embodiments, the enzyme comprising 3'-5' exonuclease activity is T4 polymerase. In particular embodiments, an enzyme which exhibits 3'-5' exonuclease enzymatic activity and is capable of primer template extension can be employed, including for example T4 or Exonucleases I, III, V. *Id.*

[00173] In some embodiments, the methods contemplated herein comprise performing sequencing and/or PCR on the 3'-5' exonuclease enzymatically processed complex discussed *supra* and elsewhere herein. In particular embodiments, a tail portion of a capture probe molecule is copied in order to generate a hybrid nucleic acid molecule. In one embodiment, the hybrid nucleic acid molecule generated comprises the target region capable of hybridizing to the capture probe module and the complement of the capture probe module tail sequence.

[00174] In a particular embodiment, genetic analysis comprises a) hybridizing one or more capture probe modules to one or more target genetic loci in a plurality of genomic DNA library clones to form one or more capture probe module-DNA library clone complexes; b) isolating the one or more capture probe module-DNA library clone complexes from a); c) enzymatically processing the one or more isolated capture probe module-DNA library clone complexes from step b); d) performing PCR on the enzymatically processed complex from c) wherein the tail portion of the capture probe molecule is copied in order to generate amplified hybrid nucleic acid molecules, wherein the amplified hybrid nucleic acid

molecules comprise a target sequence in the target genomic locus capable of hybridizing to the capture probe and the complement of the capture probe module tail sequence; and e) performing quantitative genetic analysis on the amplified hybrid nucleic acid molecules from d).

[00175] In a particular embodiment, methods for determining copy number of a specific target genetic locus are contemplated comprising: a) hybridizing one or more capture probe modules to one or more target genetic loci in a plurality of DNA library clones to form one or more capture probe module-DNA library clone complexes; b) isolating the one or more capture probe module-DNA library clone complexes from a); c) enzymatically processing the one or more isolated capture probe module-DNA library clone complexes from step b); d) performing PCR on the enzymatically processed complex from c) wherein the tail portion of the capture probe molecule is copied in order to generate amplified hybrid nucleic acid molecules, wherein the amplified hybrid nucleic acid molecules comprise a target sequence in the target genetic locus capable of hybridizing to the capture probe and the complement of the capture probe module tail sequence; e) performing PCR amplification of the amplified hybrid nucleic acid molecules in d); and f) quantitating the PCR reaction in e), wherein the quantitation allows for a determination of copy number of the specific target region.

[00176] In one embodiment, the enzymatic processing of step c) comprises performing 3'-5' exonuclease enzymatic processing on the one or more capture probe module-DNA library clone complexes from b) using an enzyme with 3'-5' exonuclease activity to remove the single stranded 3' ends; creating one or more hybrid capture probe module-cfDNA library clone molecules through the concerted action of a 5' FLAP endonuclease, DNA polymerization and nick closure by a DNA ligase; or performing 5'-3' DNA polymerase extension of the capture probe using the isolated DNA clone in the complex as a template.

[00177] In one embodiment, the enzymatic processing of step c) comprises performing 5'-3' DNA polymerase extension of the capture probe using the isolated DNA clone in the complex as a template.

[00178] In particular embodiments, PCR can be performed using any standard PCR reaction conditions well known to those of skill in the art. In certain embodiments, the PCR reaction in e) employs two PCR primers. In one embodiment, the PCR reaction in e) employs a first PCR primer that hybridizes to a repeat within the target genetic locus. In a

particular embodiment, the PCR reaction in e) employs a second PCR primer that hybridizes to the hybrid nucleic acid molecules at the target genetic locus/tail junction. In certain embodiments, the PCR reaction in e) employs a first PCR primer that hybridizes to the target genetic locus and a second PCR primer hybridizes to the amplified hybrid nucleic acid molecules at the target genetic locus/tail junction. In particular embodiments, the second primer hybridizes to the target genetic locus/tail junction such that at least one or more nucleotides of the primer hybridize to the target genetic locus and at least one or more nucleotides of the primer hybridize to the tail sequence.

[00179] In certain embodiments, the amplified hybrid nucleic acid molecules obtained from step e) are sequenced and the sequences aligned horizontally, *i.e.*, aligned to one another but not aligned to a reference sequence. In particular embodiments, steps a) through e) are repeated one or more times with one or more capture probe modules. The capture probe modules can be the same or different and designed to target either cfDNA strand of a target genetic locus. In some embodiments, when the capture probes are different, they hybridize at overlapping or adjacent target sequences within a target genetic locus in the tagged cfDNA clone library. In one embodiment, a high density capture probe strategy is used wherein a plurality of capture probes hybridize to a target genetic locus, and wherein each of the plurality of capture probes hybridizes to the target genetic locus within about 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 200 bp or more of any other capture probe that hybridizes to the target genetic locus in a tagged DNA clone library, including all intervening distances.

[00180] In some embodiments, the method can be performed using two capture probe modules per target genetic locus, wherein one hybridizes to the “Watson” strand (non-coding or template strand) upstream of the target region and one hybridizes to the “Crick” strand (coding or non-template strand) downstream of the target region.

[00181] In particular embodiments, the methods contemplated herein can further be performed multiple times with any number of capture probe modules, for example 2, 3, 4, 5, 6, 7, 8, 9, or 10 or more capture probe modules per target genetic locus any number of which hybridize to the Watson or Crick strand in any combination. In some embodiments, the sequences obtained can be aligned to one another in order to identify any of a number of differences.

[00182] In certain embodiments, a plurality of target genetic loci are interrogated, *e.g.*, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, 3000,

3500, 4000, 4500, 5000, 10000, 50000, 100000, 500000 or more in a single reaction, using one or more capture probe modules.

(b) Sequencing

[00183] In particular embodiments, the quantitative genetic analysis comprises sequencing a plurality of hybrid nucleic acid molecules, as discussed elsewhere herein, *supra*, to generate sufficient sequencing depths to obtain a plurality of unique sequencing reads. The terms “unique reads” or “unique genomic sequences” (UGS) are used interchangeably herein and are identified by grouping individual redundant reads together into a “family.” Redundant reads are sequence reads that share an identical UMIE (*e.g.*, share the same read code and the same DNA sequence start position within genomic sequence) and are derived from a single attachment event and are therefore amplification-derived “siblings” of one another. A single consensus representative of a family of redundant reads is carried forward as a unique read or UGS. Each unique read or UGS is considered a unique attachment event. The sum of unique reads corresponding to a particular capture probe is referred to as the “raw genomic depth” (RGD) for that particular capture probe. Each capture probe yields a set of unique reads that are computationally distilled from total reads by grouping into families. The unique reads for a given sample (*e.g.*, raw genomic depth for a sample) are then computed as the average of all the unique reads observed on a probe-by-probe basis. Unique reads are important because each unique read must be derived from a unique genomic DNA clone. Each unique read represents the input and analysis of a haploid equivalent of genomic DNA. The sum of unique reads is the sum of haploid genomes analyzed. The number of genomes analyzed, in turn, defines the sensitivity of the sequencing assay. By way of a non-limiting example, if the average unique read count is 100 genome equivalents, then that particular assay has a sensitivity of being able to detect one mutant read in 100, or 1%. Any observation less than this is not defensible.

[00184] Cases where there is an obvious copy number change (*e.g.*, instances of noisy probes) are excluded from the data set used to compute the sample average. Herein, a “noisy probe” refers to a probe that captures a highly variable number of unique reads among a large set identical samples (*e.g.*, a highly variable number of unique reads among 12 – 16 sample replicates). In some embodiments, the number of unique reads associated with a noisy probe is increased compared to the average number of unique reads for the sample by

50% or more. In some embodiments, the number of unique reads associated with a noisy probe is decreased compared to the average number of unique reads for the sample by 50% or more. In some embodiments, about 2% to about 4% of probes used in a particular analysis are identified as noisy probes and are excluded from calculations to determine the average number of unique reads for a given sample.

[00185] In some embodiments, sequencing reads are identified as either “on-target reads” or “off-target reads.” On-target reads possess a genomic DNA sequence that maps within the vicinity of a capture probe used to create the genomic library. In some embodiments, where each genomic sequence is physically linked to a specific capture probe and where the sequence of the genomic segment and capture probe are both determined as a unified piece of information, an on-target read is defined as any genomic sequence whose starting coordinate maps within 400 bp, and more generally within 200 bp of the 3’ end of the corresponding capture probe. Off-target reads are defined as having genomic sequence that aligns to the reference genome at a location ≥ 500 base pairs (and more often mapping to entirely different chromosomes) relative to the capture probe.

[00186] In particular embodiments, the quantitative genetic analysis comprises multiplex sequencing of hybrid nucleic acid molecules derived from a plurality of samples.

[00187] In various embodiments, the quantitative genetic analysis comprises obtaining one or more or a plurality of tagged DNA library clones, each clone comprising a first DNA sequence and a second DNA sequence, wherein the first DNA sequence comprises a sequence in a targeted genetic locus and the second DNA sequence comprises a capture probe sequence; performing a paired end sequencing reaction on the one or more clones and obtaining one or more sequencing reads or performing a sequencing reaction on the one or more clones in which a single long sequencing read of greater than about 100, 200, 300, 400, 500 or more nucleotides is obtained, wherein the read is sufficient to identify both the first DNA sequence and the second DNA sequence; and ordering or clustering the sequencing reads of the one or more clones according to the probe sequences of the sequencing reads.

(c) Bioinformatics Analysis

[00188] In various embodiments, the quantitative genetic analysis further comprises bioinformatic analysis of the sequencing reads. Bioinformatic analysis excludes any purely mental analysis performed in the absence of a composition or method for sequencing. In certain embodiments, bioinformatics analysis includes, but is not limited to:

sequence alignments; genome equivalents analysis; single nucleotide variant (SNV) analysis; gene copy number variation (CNV) analysis; measurement of chromosomal copy number; and detection of genetic lesions. In particular embodiments, bioinformatics analysis is useful to quantify the number of genome equivalents analyzed in the cfDNA clone library; to detect the genetic state of a target genetic locus; to detect genetic lesions in a target genetic locus; and to measure copy number fluctuations within a target genetic locus.

[00189] Sequence alignments may be performed between the sequence reads and one or more human reference DNA sequences. In particular embodiments, sequencing alignments can be used to detect genetic lesions in a target genetic locus including, but not limited to detection of a nucleotide transition or transversion, a nucleotide insertion or deletion, a genomic rearrangement, a change in copy number, or a gene fusion. Detection of genetic lesions that are causal or prognostic indicators may be useful in the diagnosis, prognosis, treatment, and/or monitoring of a particular genetic condition or disease.

[00190] Also contemplated herein, are methods for sequence alignment analysis that can be performed without the need for alignment to a reference sequence, referred to herein as horizontal sequence analysis. Such analysis can be performed on any sequences generated by the methods contemplated herein or any other methods. In particular embodiments, the sequence analysis comprises performing sequence alignments on the reads obtained by the methods contemplated herein.

[00191] In one embodiment, the genome equivalents in a cfDNA clone library are determined using bioinformatics-based counting after sequencing is performed. Each sequencing read is associated with a particular capture probe, and the collection of reads assigned to each capture probe is parsed into groups. Within a group, sets of individual reads share the same read code and the same DNA sequence start position within genomic sequence. These individual reads are grouped into a “family” and a single consensus representative of this family is carried forward as a “unique read.” All of the individual reads that constituted a family are derived from a single attachment event and thus, they are amplification-derived “siblings” of one another. Each unique read is considered a unique attachment event and the sum of unique reads is considered equivalent to the number of genome equivalents analyzed.

[00192] As the number of unique clones approaches the total number of possible sequence combinations, probability dictates that the same code and start site combinations will be created by independent events and that these independent events will be

inappropriately grouped within single families. The net result will be an underestimate of genome equivalents analyzed, and rare mutant reads may be discarded as sequencing errors because they overlap with wild-type reads bearing the same identifiers.

[00193] In particular embodiments, to provide an accurate analysis for cfDNA clone libraries, the number of genome equivalents analyzed is about 1/10, about 1/12, about 1/14, about 1/16, about 1/18, about 1/20, about 1/25 or less the number of possible unique clones. It should be understood that the procedure outlined above is merely illustrative and not limiting.

[00194] In some embodiments, the number of genome equivalents to be analyzed may need to be increased. To expand the depth of genome equivalents, at least two solutions are contemplated. The first solution is to use more than one adaptor set per sample. By combining adaptors, it is possible to multiplicatively expand the total number of possible clones and therefore, expand the comfortable limits of genomic input. The second solution is to expand the read code by 1, 2, 3, 4, or 5, or more bases. The number of possible read codes that differ by at least 2 bases from every other read code scales as $4^{(n-1)}$ where n is the number of bases within a read code. Thus, in a non-limiting example, if a read code is 5 nucleotides and $4^{(5-1)} = 256$; therefore, the inclusion of additional bases expands the available repertoire by a factor of four for each additional base.

[00195] In one embodiment, quantitative genetic analysis comprises bioinformatic analysis of sequencing reads to identify rare single nucleotide variants (SNV).

[00196] Next-generation sequencing has an inherent error rate of roughly 0.02-0.02%, meaning that anywhere from 1/200 to 1/500 base calls are incorrect. To detect variants and other mutations that occur at frequencies lower than this, for example at frequencies of 1 per 1000 sequences, it is necessary to invoke molecular annotation strategies. By way of a non-limiting example, analysis of 5000 unique molecules using targeted sequence capture technology would generate – at sufficient sequencing depths of >50,000 reads – a collection of 5000 unique reads, with each unique read belonging to a “family” of reads that all possess the same read code. A SNV that occurs within a family is a candidate for being a rare variant. When this same variant is observed in more than one family, it becomes a very strong candidate for being a rare variant that exists within the starting sample. In contrast, variants that occur sporadically within families are likely to be sequencing errors and variants that occur within one and only one family are either rare or the

result of a base alteration that occurred *ex vivo* (e.g., oxidation of a DNA base or PCR-introduced errors).

[00197] In one embodiment, the methods of detecting SNVs comprise introducing 10-fold more genomic input (genomes or genome equivalents) as the desired target sensitivity of the assay. In one non-limiting example, if the desired sensitivity is 2% (2 in 100), then the experimental target is an input of 2000 genomes.

[00198] In particular embodiments, bioinformatics analysis of sequencing data is used to detect or identify SNV associated with a genetic state, condition or disease, genetic mosaicism, fetal testing, paternity testing, predicting response to drug treatment, diagnosing or monitoring a medical condition, microbiome profiling, pathogen screening, and monitoring organ transplants.

[00199] In various embodiments, a method for copy number determination analysis is provided comprising obtaining one or more or a plurality of clones, each clone comprising a first DNA sequence and a second DNA sequence, wherein the first DNA sequence comprises a sequence in a targeted genetic locus and the second DNA sequence comprises a capture probe sequence. In related embodiments, a paired end sequencing reaction on the one or more clones is performed and one or more sequencing reads are obtained. In another embodiment, a sequencing reaction on the one or more clones is performed in which a single long sequencing read of greater than about 100 nucleotides is obtained, wherein the read is sufficient to identify both the first DNA sequence and the second DNA sequence. The sequencing reads of the one or more clones can be ordered or clustered according to the probe sequence of the sequencing reads.

[00200] Copy number analyses include, but are not limited to, analyses that examine the number of copies of a particular gene or mutation that occurs in a given genomic DNA sample and can further include quantitative determination of the number of copies of a given gene or sequence differences in a given sample. In particular embodiments, copy number analysis is used to detect or identify gene amplification associated with genetic states, conditions, or diseases, fetal testing, genetic mosaicism, paternity testing, predicting response to drug treatment, diagnosing or monitoring a medical condition, microbiome profiling, pathogen screening, and monitoring organ transplants.

[00201] In some embodiments, copy number analysis is used to measure chromosomal instability. In such embodiments, sets of capture probes that comprise chromosomal stability probes are used to determine copy number variations at a uniform

density across all sets of chromosomes. Copy number analyses are performed for each chromosomal stability probe and the chromosomal stability probes are then ordered according to their chromosomal target. This allows for visualization of copy number losses or gains across the genome and can serve as a measure of chromosomal stability.

[00202] In particular embodiments, bioinformatics analysis of sequencing data is used to detect or identify one or more sequences or genetic lesions in a target locus including, but not limited to detection of a nucleotide transition or transversion, a nucleotide insertion or deletion, a genomic rearrangement, a change in copy number, or a gene fusion. Detection of genetic lesions that are causal or prognostic indicators may be useful in the diagnosis, prognosis, treatment, and/or monitoring of a particular genetic condition or disease. In one embodiment, genetic lesions are associated with genetic states, conditions, or diseases, fetal testing, genetic mosaicism, paternity testing, predicting response to drug treatment, diagnosing or monitoring a medical condition, microbiome profiling, pathogen screening, and monitoring organ transplants.

D. CLINICAL APPLICATIONS OF QUANTITATIVE CNL ASSAYS

[00203] In various embodiments, the present invention contemplates a method of detecting, identifying, predicting, diagnosing, or monitoring a condition or disease in a subject by detecting a mutational change, SNP, translocation, inversion, deletion, change in copy number or other genetic variation in a region of interest.

E. CLINICAL APPLICATIONS OF QUANTITATIVE GENETIC ANALYSIS

[00204] In various embodiments, the present invention contemplates a method of detecting, identifying, predicting, diagnosing, or monitoring a condition or disease in a subject.

[00205] In particular embodiments, a method of detecting, identifying, predicting, diagnosing, or monitoring a genetic state, condition or disease in a subject comprises performing a quantitative genetic analysis of one or more target genetic loci in a DNA clone library to detect or identify a change in the sequence at the one or more target genetic loci. In some embodiments, the change is a change in copy number.

[00206] In one embodiment, a method of detecting, identifying, predicting, diagnosing, or monitoring a genetic state, condition or disease comprises isolating or obtaining cellular DNA or cfDNA from a biological sample of a subject; treating the cellular DNA or cfDNA with one or more end-repair enzymes to generate end-repaired DNA; attaching one or more adaptors to each end of the end-repaired DNA to generate a genomic DNA library; amplifying the DNA library to generate a DNA clone library; determining the number of genome equivalents in the DNA clone library; and performing a quantitative genetic analysis of one or more target genetic loci in a DNA clone library to detect or identify a change in the sequence, *e.g.*, an SNP, a translocation, an inversion, a deletion, or a change in copy number at of the one or more target genetic loci.

[00207] In particular embodiments, a method of detecting, identifying, predicting, diagnosing, or monitoring a genetic state, or genetic condition or disease selected from the group consisting of: genetic diseases; genetic mosaicism; fetal testing; paternity testing; paternity testing; predicting response to drug treatment; diagnosing or monitoring a medical condition; microbiome profiling; pathogen screening; and organ transplant monitoring comprising isolating or obtaining genomic DNA from a biological sample of a subject; treating the DNA with one or more end-repair enzymes to generate end-repaired DNA; attaching one or more adaptors to each end of the end-repaired DNA to generate a genomic DNA library; amplifying the genomic DNA library to generate a DNA clone library; determining the number of genome equivalents in the DNA clone library; and performing a quantitative genetic analysis of one or more target genetic loci in a DNA clone library to detect or identify a nucleotide transition or transversion, a nucleotide insertion or deletion, a genomic rearrangement, a change in copy number, or a gene fusion in the sequence at the one or more target genetic loci.

[00208] Illustrative examples of genetic diseases that can be detected, identified, predicted, diagnosed, or monitored with the compositions and methods contemplated herein include, but are not limited to cancer, Alzheimer's disease (APOE1), Charcot-Marie-Tooth disease, Leber hereditary optic neuropathy (LHON), Angelman syndrome (UBE3A, ubiquitin-protein ligase E3A), Prader-Willi syndrome (region in chromosome 15), β -Thalassaemia (HBB, β -Globin), Gaucher disease (type I) (GBA, Glucocerebrosidase), Cystic fibrosis (CFTR Epithelial chloride channel), Sickle cell disease (HBB, β -Globin), Tay-Sachs disease (HEXA, Hexosaminidase A), Phenylketonuria (PAH, Phenylalanine hydrolyase), Familial hypercholesterolaemia (LDLR, Low density lipoprotein

receptor), Adult polycystic kidney disease (PKD1, Polycystin), Huntington disease (HDD, Huntingtin), Neurofibromatosis type I (NF1, NF1 tumour suppressor gene), Myotonic dystrophy (DM, Myotonin), Tuberous sclerosis (TSC1, Tuberin), Achondroplasia (FGFR3, Fibroblast growth factor receptor), Fragile X syndrome (FMR1, RNA-binding protein), Duchenne muscular dystrophy (DMD, Dystrophin), Haemophilia A (F8C, Blood coagulation factor VIII), Lesch–Nyhan syndrome (HPRT1, Hypoxanthine guanine ribosyltransferase 1), and Adrenoleukodystrophy (ABCD1).

[00209] Illustrative examples of cancers that can be detected, identified, predicted, diagnosed, or monitored with the compositions and methods contemplated herein include, but are not limited to: B cell cancer, *e.g.*, multiple myeloma, melanomas, breast cancer, lung cancer (such as non-small cell lung carcinoma or NSCLC), bronchus cancer, colorectal cancer, prostate cancer, pancreatic cancer, stomach cancer, ovarian cancer, urinary bladder cancer, brain or central nervous system cancer, peripheral nervous system cancer, esophageal cancer, cervical cancer, uterine or endometrial cancer, cancer of the oral cavity or pharynx, liver cancer, kidney cancer, testicular cancer, biliary tract cancer, small bowel or appendix cancer, salivary gland cancer, thyroid gland cancer, adrenal gland cancer, osteosarcoma, chondrosarcoma, cancer of hematological tissues, adenocarcinomas, inflammatory myofibroblastic tumors, gastrointestinal stromal tumor (GIST), colon cancer, multiple myeloma (MM), myelodysplastic syndrome (MDS), myeloproliferative disorder (MPD), acute lymphocytic leukemia (ALL), acute myelocytic leukemia (AML), chronic myelocytic leukemia (CML), chronic lymphocytic leukemia (CLL), polycythemia Vera, Hodgkin lymphoma, non-Hodgkin lymphoma (NHL), soft-tissue sarcoma, fibrosarcoma, myxosarcoma, liposarcoma, osteogenic sarcoma, chordoma, angiosarcoma, endotheliosarcoma, lymphangiosarcoma, lymphangioendotheliosarcoma, synovioma, mesothelioma, Ewing's tumor, leiomyosarcoma, rhabdomyosarcoma, squamous cell carcinoma, basal cell carcinoma, adenocarcinoma, sweat gland carcinoma, sebaceous gland carcinoma, papillary carcinoma, papillary adenocarcinomas, medullary carcinoma, bronchogenic carcinoma, renal cell carcinoma, hepatoma, bile duct carcinoma, choriocarcinoma, seminoma, embryonal carcinoma, Wilms' tumor, bladder carcinoma, epithelial carcinoma, glioma, astrocytoma, medulloblastoma, craniopharyngioma, ependymoma, pinealoma, hemangioblastoma, acoustic neuroma, oligodendroglioma, meningioma, neuroblastoma, retinoblastoma, follicular lymphoma, diffuse large B-cell lymphoma, mantle cell lymphoma, hepatocellular carcinoma, thyroid cancer, gastric cancer,

head and neck cancer, small cell cancers, essential thrombocythemia, agnogenic myeloid metaplasia, hypereosinophilic syndrome, systemic mastocytosis, familial hypereosinophilia, chronic eosinophilic leukemia, neuroendocrine cancers, carcinoid tumors, and the like.

[00210] In one embodiment, the genetic lesion is a lesion annotated in the Cosmic database (the lesions and sequence data are available online and can be downloaded from the Cancer Gene Census section of the Cosmic website) or a lesion annotated in the Cancer Genome Atlas (the lesions and sequence data are available online and can be downloaded from The Cancer Genome Atlas website).

[00211] Illustrative examples of genes that harbor one or more genetic lesions associated with cancer that can be detected, identified, predicted, diagnosed, or monitored with the compositions and methods contemplated herein include, but are not limited to ABCB1, ABCC2, ABCC4, ABCG2, ABL1, ABL2, AKT1, AKT2, AKT3, ALDH4A1, ALK, APC, AR, ARAF, ARFRP1, ARID1A, ATM, ATR, AURKA, AURKB, BCL2, BCL2A1, BCL2L1, BCL2L2, BCL6, BRAF, BRCA1, BRCA2, Clorf144, CARD11, CBL, CCND1, CCND2, CCND3, CCNE1, CDH1, CDH2, CDH20, CDH5, CDK4, CDK6, CDK8, CDKN2A, CDKN2B, CDKN2C, CEBPA, CHEK1, CHEK2, CRKL, CRLF2, CTNNB1, CYP1B1, CYP2C19, CYP2C8, CYP2D6, CYP3A4, CYP3A5, DNMT3A, DOT1L, DPYD, EGFR, EPHA3, EPHA5, EPHA6, EPHA7, EPHB1, EPHB4, EPHB6, EPHX1, ERBB2, ERBB3, ERBB4, ERCC2, ERG, ESR1, ESR2, ETV1, ETV4, ETV5, ETV6, EWSR1, EZH2, FANCA, FBXW7, FCGR3A, FGFR1, FGFR2, FGFR3, FGFR4, FLT1, FLT3, FLT4, FOXP4, GATA1, GNA11, GNAQ, GNAS, GPR124, GSTP1, GUCY1A2, HOXA3, HRAS, HSP90AA1, IDH1, IDH2, IGF1R, IGF2R, IKBKE, IKZF1, INHBA, IRS2, ITPA, JAK1, JAK2, JAK3, JUN, KDR, KIT, KRAS, LRP1B, LRP2, LTK, MAN1B1, MAP2K1, MAP2K2, MAP2K4, MCL1, MDM2, MDM4, MEN1, MET, MITF, MLH1, MLL, MPL, MRE11A, MSH2, MSH6, MTHFR, MTOR, MUTYH, MYC, MYCL1, MYCN, NF1, NF2, NKX2-1, NOTCH1, NPM1, NQO1, NRAS, NRP2, NTRK1, NTRK3, PAK3, PAX5, PDGFRA, PDGFRB, PIK3CA, PIK3R1, PKHD1, PLCG1, PRKDC, PTCH1, PTEN, PTPN11, PTPRD, RAF1, RARA, RB1, RET, RICTOR, RPTOR, RUNX1, SLC19A1, SLC22A2, SLC01B3, SMAD2, SMAD3, SMAD4, SMARCA4, SMARCB1, SMO, SOD2, SOX10, SOX2, SRC, STK11, SULT1A1, TBX22, TET2, TGFB2, TMPRSS2, TNFRSF14, TOP1, TP53, TPMT, TSC1, TSC2, TYMS, UGT1A1, UMPS, USP9X, VHL, and WT1.

[00212] In particular embodiments, the genetic lesion comprises a nucleotide transition or transversion, a nucleotide insertion or deletion, a genomic rearrangement, a change in copy number, or a gene fusion.

[00213] In one embodiment, the genetic lesion is a gene fusion that fuses the 3' coding region of the ALK gene to another gene.

[00214] In one embodiment, the genetic lesion is a gene fusion that fuses the 3' coding region of the ALK gene to the EML4 gene.

[00215] Illustrative examples of conditions suitable for fetal testing that can be detected, identified, predicted, diagnosed, or monitored with the compositions and methods contemplated herein include but are not limited to: Down Syndrome (Trisomy 21), Edwards Syndrome (Trisomy 18), Patau Syndrome (Trisomy 13), Klinefelter's Syndrome (XXY), Triple X syndrome, XYY syndrome, Trisomy 8, Trisomy 16, Turner Syndrome (XO), Robertsonian translocation, DiGeorge Syndrome and Wolf-Hirschhorn Syndrome.

[00216] Illustrative examples of alleles suitable for paternity testing that can be detected, identified, predicted, diagnosed, or monitored with the compositions and methods contemplated herein include but are not limited to 16 or more of: D20S1082, D6S474, D12ATA63, D22S1045, D10S1248, D1S1677, D11S4463, D4S2364, D9S1122, D2S1776, D10S1425, D3S3053, D5S2500, D1S1627, D3S4529, D2S441, D17S974, D6S1017, D4S2408, D9S2157, Amelogenin, D17S1301, D1GATA113, D18S853, D20S482, and D14S1434.

[00217] Illustrative examples of genes suitable for predicting the response to drug treatment that can be detected, identified, predicted, diagnosed, or monitored with the compositions and methods contemplated herein include, but are not limited to, one or more of the following genes: ABCB1 (ATP-binding cassette, sub-family B (MDR/TAP), member 1), ACE (angiotensin I converting enzyme), ADH1A (alcohol dehydrogenase 1A (class I), alpha polypeptide), ADH1B (alcohol dehydrogenase 1B (class I), beta polypeptide), ADH1C (alcohol dehydrogenase 1C (class I), gamma polypeptide), ADRB1 (adrenergic, beta-1-, receptor), ADRB2 (adrenergic, beta-2-, receptor, surface), AHR (aryl hydrocarbon receptor), ALDH1A1 (aldehyde dehydrogenase 1 family, member A1), ALOX5 (arachidonate 5-lipoxygenase), BRCA1 (breast cancer 1, early onset), COMT (catechol-O-methyltransferase), CYP2A6 (cytochrome P450, family 2, subfamily A, polypeptide 6), CYP2B6 (cytochrome P450, family 2, subfamily B, polypeptide 6), CYP2C9 (cytochrome P450, family 2, subfamily C, polypeptide 9), CYP2C19 (cytochrome P450, family 2, subfamily C,

polypeptide 19), CYP2D6 (cytochrome P450, family 2, subfamily D, polypeptide 6), CYP2J2 (cytochrome P450, family 2, subfamily J, polypeptide 2), CYP3A4 (cytochrome P450, family 3, subfamily A, polypeptide 4), CYP3A5 (cytochrome P450, family 3, subfamily A, polypeptide 5), DPYD (dihydropyrimidine dehydrogenase), DRD2 (dopamine receptor D2), F5 (coagulation factor V), GSTP1 (glutathione S-transferase pi), HMGCR (3-hydroxy-3-methylglutaryl-Coenzyme A reductase), KCNH2 (potassium voltage-gated channel, subfamily H (eag-related), member 2), KCNJ11 (potassium inwardly-rectifying channel, subfamily J, member 11), MTHFR (5,10-methylenetetrahydrofolate reductase (NADPH)), NQO1 (NAD(P)H dehydrogenase, quinone 1), P2RY1 (purinergic receptor P2Y, G-protein coupled, 1), P2RY12 (purinergic receptor P2Y, G-protein coupled, 12), PTGIS (prostaglandin I₂ (prostacyclin) synthase), SCN5A (sodium channel, voltage-gated, type V, alpha (long QT syndrome 3)), SLC19A1 (solute carrier family 19 (folate transporter), member 1), SLCO1B1 (solute carrier organic anion transporter family, member 1B1), SULT1A1 (sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1), TPMT (thiopurine S-methyltransferase), TYMS (thymidylate synthetase), UGT1A1 (UDP glucuronosyltransferase 1 family, polypeptide A1), VDR (vitamin D (1,25- dihydroxyvitamin D₃) receptor), VKORC1 (vitamin K epoxide reductase complex, subunit 1).

[00218] Illustrative examples of medical conditions that can be detected, identified, predicted, diagnosed, or monitored with the compositions and methods contemplated herein include, but are not limited to: stroke, transient ischemic attack, traumatic brain injury, heart disease, heart attack, angina, atherosclerosis, and high blood pressure.

[00219] Illustrative examples of pathogens that can be screened for with the compositions and methods contemplated herein include, but are not limited to: bacteria fungi, and viruses.

[00220] Illustrative examples of bacterial species that can be screened for with the compositions and methods contemplated herein include, but are not limited to: a *Mycobacterium* spp., a *Pneumococcus* spp., an *Escherichia* spp., a *Campylobacter* spp., a *Corynebacterium* spp., a *Clostridium* spp., a *Streptococcus* spp., a *Staphylococcus* spp., a *Pseudomonas* spp., a *Shigella* spp., a *Treponema* spp., or a *Salmonella* spp.

[00221] Illustrative examples of fungal species that can be screened for with the compositions and methods contemplated herein include, but are not limited to: an *Aspergillus* spp., a *Blastomyces* spp., a *Candida* spp., a *Coccidioides* spp., a *Cryptococcus*

spp., dermatophytes, a *Tinea* spp., a *Trichophyton* spp., a *Microsporum* spp., a *Fusarium* spp., a *Histoplasma* spp., a *Mucoromycotina* spp., a *Pneumocystis* spp., a *Sporothrix* spp., an *Exserophilum* spp., or a *Cladosporium* spp.

[00222] Illustrative examples of viruses that can be screened for with the compositions and methods contemplated herein include, but are not limited to: Influenza A such as H1N1, H1N2, H3N2 and H5N1 (bird flu), Influenza B, Influenza C virus, Hepatitis A virus, Hepatitis B virus, Hepatitis C virus, Hepatitis D virus, Hepatitis E virus, Rotavirus, any virus of the Norwalk virus group, enteric adenoviruses, parvovirus, Dengue fever virus, Monkey pox, Mononegavirales, Lyssavirus such as rabies virus, Lagos bat virus, Mokola virus, Duvenhage virus, European bat virus 1 & 2 and Australian bat virus, Ephemerovirus, Vesiculovirus, Vesicular Stomatitis Virus (VSV), Herpesviruses such as Herpes simplex virus types 1 and 2, varicella zoster, cytomegalovirus, Epstein-Bar virus (EBV), human herpesviruses (HHV), human herpesvirus type 6 and 8, Moloney murine leukemia virus (M-MuLV), Moloney murine sarcoma virus (MoMSV), Harvey murine sarcoma virus (HaMuSV), murine mammary tumor virus (MuMTV), gibbon ape leukemia virus (GaLV), feline leukemia virus (FLV), spumavirus, Friend murine leukemia virus, Murine Stem Cell Virus (MSCV) and Rous Sarcoma Virus (RSV), HIV (human immunodeficiency virus; including HIV type 1, and HIV type 2), visna-maedi virus (VMV) virus, the caprine arthritis-encephalitis virus (CAEV), equine infectious anemia virus (EIAV), feline immunodeficiency virus (FIV), bovine immune deficiency virus (BIV), and simian immunodeficiency virus (SIV), papilloma virus, murine gammaherpesvirus, Arenaviruses such as Argentine hemorrhagic fever virus, Bolivian hemorrhagic fever virus, Sabia-associated hemorrhagic fever virus, Venezuelan hemorrhagic fever virus, Lassa fever virus, Machupo virus, Lymphocytic choriomeningitis virus (LCMV), Bunyaviridae such as Crimean-Congo hemorrhagic fever virus, Hantavirus, hemorrhagic fever with renal syndrome causing virus, Rift Valley fever virus, Filoviridae (filovirus) including Ebola hemorrhagic fever and Marburg hemorrhagic fever, Flaviviridae including Kaysanur Forest disease virus, Omsk hemorrhagic fever virus, Tick-borne encephalitis causing virus and Paramyxoviridae such as Hendra virus and Nipah virus, variola major and variola minor (smallpox), alphaviruses such as Venezuelan equine encephalitis virus, eastern equine encephalitis virus, western equine encephalitis virus, SARS-associated coronavirus (SARS-CoV), West Nile virus, and any encephalitis causing virus.

[00223] Illustrative examples of genes suitable for monitoring an organ transplant in a transplant recipient that can be detected, identified, predicted, diagnosed, or monitored with the compositions and methods contemplated herein include, but are not limited to, one or more of the following genes: HLA-A, HLA-B, HLA-C, HLA-DR, HLA-DP, and HLA-DQ.

[00224] In particular embodiments, a bioinformatic analysis is used to quantify the number of genome equivalents analyzed in the cfDNA clone library; detect genetic variants in a target genetic locus; detect mutations within a target genetic locus; detect genetic fusions within a target genetic locus; or measure copy number fluctuations within a target genetic locus.

F. COMPANION DIAGNOSTICS

[00225] In various embodiments, a companion diagnostic for a genetic disease is provided, comprising: isolating or obtaining genomic DNA from a biological sample of a subject; treating the DNA with one or more end-repair enzymes to generate end-repaired DNA; attaching one or more adaptors to each end of the end-repaired DNA to generate a DNA library; amplifying the DNA library to generate a DNA clone library; determining the number of genome equivalents in the DNA clone library; and performing a quantitative genetic analysis of one or more biomarkers associated with the genetic disease in the DNA clone library, wherein detection of, or failure to detect, at least one of the one or more biomarkers indicates whether the subject should be treated for the genetic disease. In some embodiments, the DNA is cfDNA. In particular embodiments, the DNA is cellular DNA.

[00226] As used herein, the term “companion diagnostic” refers to a diagnostic test that is linked to a particular anti-cancer therapy. In a particular embodiment, the diagnostic methods comprise detection of genetic lesion in a biomarker associated with in a biological sample, thereby allowing for prompt identification of patients should or should not be treated with the anti-cancer therapy.

[00227] Anti-cancer therapy includes, but is not limited to surgery, radiation, chemotherapeutics, anti-cancer drugs, and immunomodulators.

[00228] Illustrative examples of anti-cancer drugs include, but are not limited to: alkylating agents such as thiotepa and cyclophosphamide (CYTOXAN™); alkyl sulfonates such as busulfan, improsulfan and piposulfan; aziridines such as benzodopa,

carboquone, meturedopa, and uredopa; ethylenimines and methylamelamines including altretamine, triethylenemelamine, trietylenephosphoramidate, triethylenethiophosphoramidate and trimethylolomelamine resins; nitrogen mustards such as chlorambucil, chlornaphazine, cholophosphamide, estramustine, ifosfamide, mechlorethamine, mechlorethamine oxide hydrochloride, melphalan, novembichin, phenesterine, prednimustine, trofosfamide, uracil mustard; nitrosureas such as carmustine, chlorozotocin, fotemustine, lomustine, nimustine, ranimustine; antibiotics such as aclacinomysins, actinomycin, authramycin, azaserine, bleomycins, cactinomycin, calicheamicin, carabycin, carminomycin, carzinophilin, chromomycins, dactinomycin, daunorubicin, detorubicin, 6-diazo-5-oxo-L-norleucine, doxorubicin and its pegylated formulations, epirubicin, esorubicin, idarubicin, marcellomycin, mitomycins, mycophenolic acid, nogalamycin, olivomycins, peplomycin, potfiromycin, puromycin, quelamycin, rodorubicin, streptonigrin, streptozocin, tubercidin, ubenimex, zinostatin, zorubicin; anti-metabolites such as methotrexate and 5-fluorouracil (5-FU); folic acid analogues such as denopterin, methotrexate, pteropterin, trimetrexate; purine analogs such as fludarabine, 6-mercaptopurine, thiamiprine, thioguanine; pyrimidine analogs such as ancitabine, azacitidine, 6-azauridine, carmofur, cytarabine, dideoxyuridine, doxifluridine, enocitabine, floxuridine, 5-FU; androgens such as calusterone, dromostanolone propionate, epitostanol, mepitiostane, testolactone; anti-adrenals such as aminoglutethimide, mitotane, trilostane; folic acid replenisher such as frolinic acid; aceglatone; aldophosphamide glycoside; aminolevulinic acid; amsacrine; bestrabucil; bisantrene; edatraxate; defofamine; demecolcine; diaziquone; elformithine; elliptinium acetate; etoglucid; gallium nitrate; hydroxyurea; lentinan; lonidamine; mitoguanzone; mitoxantrone; mopidamol; nitracrine; pentostatin; phenamet; pirarubicin; podophyllinic acid; 2-ethylhydrazide; procarbazine; PSK®; razoxane; sizofiran; spirogermanium; tenuazonic acid; triaziquone; 2, 2',2''-trichlorotriethylamine; urethan; vindesine; dacarbazine; mannomustine; mitobronitol; mitolactol; pipobroman; gacytosine; arabinoside ("Ara-C"); cyclophosphamide; thiotepa; taxoids, *e.g.*, paclitaxel (TAXOL®, Bristol-Myers Squibb Oncology, Princeton, N.J.) and doxetaxel (TAXOTERE®, Rhne-Poulenc Rorer, Antony, France); chlorambucil; gemcitabine; 6-thioguanine; mercaptopurine; methotrexate; platinum analogs such as cisplatin and carboplatin; vinblastine; platinum; etoposide (VP-16); ifosfamide; mitomycin C; mitoxantrone; vincristine; vinorelbine; navelbine; novantrone; teniposide; aminopterin; xeloda; ibandronate; CPT-11; topoisomerase inhibitor RFS 2000; difluoromethylomithine (DMFO); retinoic acid derivatives such as Targretin™ (bexarotene), Panretin™ (alitretinoin);

ONTAK™ (denileukin diftitox) ; esperamicins; capecitabine; and pharmaceutically acceptable salts, acids or derivatives of any of the above. Also included in this definition are anti-hormonal agents that act to regulate or inhibit hormone action on cancers such as anti-estrogens including for example tamoxifen, raloxifene, aromatase inhibiting 4(5)-imidazoles, 4-hydroxytamoxifen, trioxifene, keoxifene, LY117018, onapristone, and toremifene (Fareston); and anti-androgens such as flutamide, nilutamide, bicalutamide, leuprolide, and goserelin; and pharmaceutically acceptable salts, acids or derivatives of any of the above.

[00229] Illustrative examples of immunomodulators include, but are not limited to: cyclosporine, tacrolimus, tresperimus, pimecrolimus, sirolimus, verolimus, laflunimus, laquinimod and imiquimod, as well as analogs, derivatives, salts, ions and complexes thereof.

[00230] In some embodiments, an anti-cancer drug may include a poly-ADP ribose polymerase (PARP) inhibitor. Illustrative examples of PARP inhibitors include, but are not limited to, olaparib (AZD-2281), rucaparib (AG014699 or PF-01367338, niraparib (MK-4827), talazoparib (BMN-673) veliparib (ABT-888), CEP 9722, E7016, BGB-290, 3-aminobenzamide.

[00231] All publications, patent applications, and issued patents cited in this specification are herein incorporated by reference as if each individual publication, patent application, or issued patent were specifically and individually indicated to be incorporated by reference. In particular, the entire contents of International PCT Publication No. WO 2016/028316 are specifically incorporated by reference.

[00232] Although the foregoing invention has been described in some detail by way of illustration and example for purposes of clarity of understanding, it will be readily apparent to one of ordinary skill in the art in light of the teachings of this invention that certain changes and modifications may be made thereto without departing from the spirit or scope of the appended claims. The following examples are provided by way of illustration only and not by way of limitation. Those of skill in the art will readily recognize a variety of noncritical parameters that could be changed or modified to yield essentially similar results.

EXAMPLES

Example 1: Copy Number Analysis of Samples Containing Blends Of Fragmented Genomic DNA

[00233] Meticulous blends of fragmented genomic DNA were generated that contained DNA derived from Δ ATM or Δ BRCA2 immortalized human samples spiked into a

fragmented wild-type human gDNA sample. The advantage of this sample type is that the composition can be carefully controlled and sample availability is essentially unlimited.

[00234] Wild-type, human female genomic DNA was purified from whole blood samples donated by a healthy volunteer. Genomic DNA isolated from an immortalized cell harboring a heterozygous deletion covering the entire ATM gene (NA09596, Δ ATM) and a separate sample bearing a heterozygous deletion of BRCA2 (NA02718, Δ BRCA2) were obtained from the Coriell repository. Importantly, these samples appeared to have an otherwise normal ploidy across the remainder of the genomes. The Δ ATM sample was derived from a male donor and was therefore also hemizygous in copy number for the X-linked AR gene. Cell free DNA (cfDNA) was obtained from healthy donor plasma samples of female or male origin. For library construction, genomic DNA was sonicated on a setting of 200 bp with a Covaris instrument, then further size selected using a “two-sided” DNA bead purification. Library input DNA samples are shown in FIG. 7.

[00235] Appropriate combinations of fragmented and cfDNA samples were blended to defined percentages, end-repaired, and converted to genomic libraries. Approximately 500 ng of each library was combined in sets of eight samples and hybridized to the copy number loss (CNL) prostate probe pool that contained 2304 DNA probes. Following sample processing, each set of eight samples was sequenced on an Illumina NextSeq NGS instrument to a depth of ~480 million pass-filter reads; this corresponds to 60 million reads/sample. Roughly 95% of reads possessed legitimate sample ID tags and aligned to the human reference genome and of these, ~98% mapped to the intended target loci. The overall sequencing depth, measured as the number of reads per input genome per probe (calculated as on-target reads (60 million) divided by average genome depth (2500) and divided by probe count (2400)) was approximately 10 reads per genome per probe. A graphic representation of the copy number loss analysis is shown in FIG. 1. Copy number perturbations are highlighted by arrows. (Sample 1, 5% male DNA into female DNA; sample 2, 5% Δ ATM DNA (male) into female DNA; sample 3, 5% Δ BRCA2 DNA (female) into female DNA; sample 4, pure female DNA).

[00236] The CNL caller identifies redundant reads and condenses these into a single consensus reads that are then quantified at each probe location. This information was further condensed into gene-by-gene copy number averages. Finally, a statistical significance

was assigned to deviations detected in each CNL measurement; this is shown graphically as the $\log_{10}$ P-value of statistical significance.

[00237] FIG. 8 shows box-and-whisker plots of copy number determinations for the AR (FIG. 8B) and ATM (FIG. 8C) genes in fragmented and blended genomic libraries. Because the Δ ATM sample is male, the AR gene (X-linked, hemizygous) and the ATM gene both exhibited CNL behavior. As anticipated, the magnitude of measured copy variation was modest. The statistical analysis shown in FIG. 9B demonstrates that the observed copy fluctuation was statistically significant. Moreover, very little significant fluctuation was observed in the remaining genes that were predicted to exhibit uniform copy characteristics. These values correlated well with frequencies predicted for the various genomic blends. FIG. 10 shows that statistically significant copy fluctuation was also readily observed in samples that were primarily cfDNA with minor spike-ins of either cfDNA from the opposite sex or minor additions of fragmented gDNA. These values correlated well with frequencies predicted for the various genomic blends. The results seen with both fragmented gDNA and with cfDNA were comparable, thereby demonstrating the integrity of the assay and suggesting that the integrity will translate to clinical samples.

[00238] These data demonstrate the ability of the assay system to detect subtle changes in gene copy number down to minor allele frequencies of 2%. While the focus of demonstrated examples presented is on copy number loss, the technology is equally well suited to the detection of copy number gains, including increases in gene copy that occur through chromosomal arm duplications and focal amplifications. This assay further retains the ability to detect other types of genomic variants, including SNVs, indels and gene fusions (chromosomal rearrangements). Importantly, these data demonstrate that the method can be applied to genomic DNA derived from plasma, but also to genomic DNA derived from other sources such as tissue and other bodily sources.

Example 2: Copy number analysis of cfDNA from healthy donors and a cancer patient

[00239] The following example illustrate the manner in which the molecular features added during genomic library construction and post-hybridization processing are used to generate copy number analysis. DNA was extracted from the plasma of sixteen healthy donors and one castration-resistant prostate cancer patient using the Qiagen Circulating Nucleic Acids Extraction kit (Qiagen, Hilden, Germany). The yield of double-strand DNA was quantified using a Qubit fluorometer (Thermo Fisher, Waltham, MA) and

the corresponding hsDNA quantitation kit. Size analysis was performed using gel electrophoresis on 2% agarose gels with PCR markers as size standards (New England Biolabs, Ipswich, MA). Approximately 40 – 100 ng of cfDNA, depending on the yield of cfDNA from the sample, was used for library construction.

[00240] The basic features of library construction are illustrated in FIG. 11A – 11C. The cfDNA was first dephosphorylated and then repaired to blunt ends in a two-step process. Short, 10 nt anchor sequences consisting of a phosphorylated ligation strand and an inert partner strand were then ligated to the cfDNA. The eight oligonucleotides used to create the set of four anchor sequences are shown in Table 1.

Table 1: Ligation anchor oligonucleotides

Oligo ID	Nucleic Acid Sequence	SEQ ID NO:
Partner strand oLigation strand oligoo_16-1	GTATGCC[3-dA-Q]*	1
Partner strand oLigation strand oligoo_16-2	AGCGTTA[3-dC-Q]*	2
Partner strand oLigation strand oligoo_16-3	TCGACAT[3-dG-Q]*	3
Partner strand oLigation strand oligoo_16-4	CATCAGG[3-dT-Q]*	4
Ligation strand oligo_16-1	/5Phos/TGG CAT ACG T**	5
Ligation strand oligo_16-2	/5Phos/GTA ACG CTA G**	6
Ligation strand oligo_16-3	/5Phos/CAT GTC GAT C**	7
Ligation strand oligo_16-4	/5Phos/ACC TGA TGC A**	8

*[3-d(A, C, G, or T)-Q] denotes a modified base in which the hydroxyl group resides on the 2' position of the ribose ring

** /5Phos/ denotes the chemical addition of a 5' phosphate group to the 5' base position

[00241] The adaptor structures were completed by the addition of full-length adaptor sequences that annealed to the anchor sequence. Thirty-two sets of adaptor sequences, each composed of 240 members, are shown in FIG. 12 – FIG. 22. These adaptors were attached to the cfDNA and extended through the concerted actions of polynucleotide kinase, DNA polymerase and DNA ligase to generate genomic libraries. As a pre-sequencing quality control step, the resulting genomic libraries were quantified by qPCR for depth of coverage. The genomic libraries were then amplified and hybridized to probe sets targeting specific genes (FIG. 11B). Following hybridization, primer extension of the probe was used to copy the captured genomic sequences and the information encoded in the attached adaptor (FIG. 11C). An example of post sequencing analysis using standard next-generation analysis software is shown in FIG. 11D. This analysis was performed on a sequencing run that

contained 32 samples (28 cancer patient samples and 4 wild-type controls) and it displays the overall distribution of sequencing reads.

[00242] A central feature of the targeted hybrid capture platform described herein is that it provides multiple types of genomic information. One essential function of capture probes is to provide mutation detection across target regions at a high depth of coverage. This function is governed by the sequence context, density, and placement of the capture probes and is illustrated in FIG. 23 with the TP53 gene (TP53 probe sequences are shown in Table 2 below). Of equal significance, the targeted hybrid capture platform assay generated a readout of equal depth of coverage in regions where no significant mutations were detected. These data are critical to physicians and patients as they add statistical significance in cases where no deleterious mutations were detected.

Table 2: TP53 Probes

Name	Sequence	SEQ ID NO:
TP53_1	GGCACAGACCCTCTCACTCATGTGATGTCATCTCTCCTCC	7689
TP53_2	ATGGGGGTGGGAGGCTGTCAGTGGGGAACAAGAAGTGGAG	7690
TP53_3	GTCAGTCTGAGTCAGGCCCTTCTGTCTTGAACATGAGTTT	7691
TP53_4	CCTGAAGTCCAAAAAGGGTCAGTCTACCTCCCGCCATAAA	7692
TP53_5	TCATGCTGGATCCCCACTTTTCTCTTGCAGCAGCCAGAC	7693
TP53_6	GTTGGGGTGGGGGTGGTGGGCCTGCCCTTCCAATGGATCC	7694
TP53_7	CAGTTTCCATAGGTCTGAAAATGTTTCTGACTCAGAGGG	7695
TP53_8	CTGCCATGGAGGAGCCGAGTCAGATCCTAGCGTCGAGCC	7696
TP53_9	GCAGAGACCTGTGGGAAGCGAAAATTCATGGGACTGACT	7697
TP53_10	CTGGGGGGCTGGGGGGCTGAGGACCTGGTCCTCTGACTGC	7698
TP53_11	GCAGGGGGGATACGGCCAGGCATTGAAGTCTCATGGAAGCC	7699
TP53_12	GTGGCCCCTGCACCAGCAGCTCCTACACGGCGGCCCTG	7700
TP53_13	GGGGGGAGCAGCCTCTGGCATTCTGGGAGCTTCATCTGGA	7701
TP53_14	CCGTGCAAGTCACAGACTTGGCTGTCCAGAATGCAAGAA	7702
TP53_15	CCCCGGACGATATTGAACAATGGTTCACTGAAGACCCAGG	7703
TP53_16	CCAGAAAACCTACCAGGGCAGCTACGGTTTCCGTCTGGGC	7704
TP53_17	TAGGTTTTCTGGGAAGGGACAGAAGATGACAGGGGCCAGG	7705
TP53_18	TGCTTTATCTGTTCACTTGTGCCCTGACTTTCAACTCTGT	7706
TP53_19	CCTGGGCAACCAGCCCTGTCGTCTCTCCAGCCCCAGCTGC	7707
TP53_20	TTTGCCAACTGGCCAAGACCTGCCCTGTGCAGCTGTGGGT	7708
TP53_21	CCATCGCTATCTGAGCAGCGCTCATGGTGGGGGCAGCGCC	7709
TP53_22	GCCATCTACAAGCAGTCACAGCACATGACGGAGGTTGTGA	7710

TP53_23	CATGGCGCGGACGCGGGTGCCGGGCGGGGGTGTGGAATCA	7711
TP53_24	CCAGGGTCCCCAGGCCTCTGATTCTCACTGATTGCTCTT	7712
TP53_25	GAGGGCCACTGACAACCACCCTTAACCCCTCCTCCCAGAG	7713
TP53_26	CCTCAGGCGGCTCATAGGGCACCACCACACTATGTCGAAA	7714
TP53_27	AGGAAATTTGCGTGTGGAGTATTTGGATGACAGAAACACT	7715
TP53_28	CTTGCCACAGGTCTCCCCAAGGCGCACTGGCCTCATCTTG	7716
TP53_29	GAGGCAAGCAGAGGCTGGGGCACAGCAGGCCAGTGTGCAG	7717
TP53_30	CCTGGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTC	7718
TP53_31	ACTACATGTGTAACAGTTCCTGCATGGGCGGCATGAACCG	7719
TP53_32	GGACAGGTAGGACCTGATTTCCCTTACTGCCTCTTGCTTCT	7720
TP53_33	CTGCACCCTTGGTCTCCTCCACCGCTTCTTGTCTGCTTG	7721
TP53_34	TCTCTTTTTCCTATCCTGAGTAGTGGTAATCTACTGGGACG	7722
TP53_35	CCTCGCTTAGTGCTCCCTGGGGGCAGCTCGTGGTGAGGCT	7723
TP53_36	GACCGGCGCACAGAGGAAGAGAATCTCCGCAAGAAAGGGG	7724
TP53_37	TCTCCAGGACAGGCACAAACACGCACCTCAAAGCTGTTC	7725
TP53_38	TGCCTCAGATTCACCTTTTATCACCTTTTCTTGCTCTTTC	7726
TP53_39	GGCATTTTGTAGTGTTAGACTGGAACTTTCCACTTGATAA	7727
TP53_40	CCTGAAGGGTGAAATATTCTCCATCCAGTGGTTTCTTCTT	7728
TP53_41	CCTAGCACTGCCCAACAACACCAGCTCCTCTCCCCAGCCA	7729
TP53_42	CATCTTTTAACTCAGGTACTGTGTATATACTTACTTCTCC	7730
TP53_43	ATGGCTTTCCAACCTAGGAAGGCAGGGGAGTAGGGCCAGG	7731
TP53_44	CCTGGAGTGAGCCCTGCTCCCCCTGGCTCCTTCCCAGCC	7732
TP53_45	TCCGAGAGCTGAATGAGGCCTTGGAAGTCAAGGATGCCCA	7733

[00243] The linkage of the capture probe with captured genomic sequence (FIG. 11C) also facilitated measurement of genomic depth at each probe location. The number of unique reads associated with every capture probe used in the experiment was measured (FIG. 24). The data shown in FIG. 24 was derived from a sequencing run in which 16 healthy donor cfDNA samples were analyzed. The depth of unique reads encountered in each sample at one probe location in the TP53 gene were calculated (Raw unique read counts shown in FIG. 24A). Each sample comprised a unique library depth, as reflected in the broad sample-to-sample distribution of unique reads. The global average of unique read depth across all 2596 capture probes in the experiment was also calculated (FIG. 24B). Significantly, normalization of the observed read depth at the single probe site displayed in FIG. 24C by the global unique read depth measured for all probes revealed a uniform density of normalized unique reads. These data indicate that the capture performance of a particular

probe chosen for analysis was uniform from sample-to-sample and proportional to the genomic depth of each individual library.

[00244] This same normalization function was applied to the 45 TP53-specific probes shown in FIG. 23 (normalization data shown in FIG. 25). Whereas FIG. 23 shows the aggregate contribution of all probes to the sequencing depth of TP53 coding regions, FIG. 25 shows the normalized depth retrieved by each individual probe. The normalized depth retrieved by each individual probe was generally consistent from sample-to-sample for any given probe but somewhat variable when one probe was compared to another. Several factors governed the differences in the post-normalization capture depths observed between probes, the most significant being the placement of probes relative to one another and the proximity of probes to genomic repeat regions. Not all probes exhibited uniform capture behavior; two probes whose capture performance were not consistent are highlighted by arrows in FIG. 25. However, these data indicate that such probes are rare and easily identified. As such, and they can be excluded from downstream copy number analysis.

[00245] The uniform capture performance exhibited by the 45 TP53 targeting probes in FIG. 25 is a general feature of the targeted hybrid capture platform described herein. In FIG. 26, the average capture depth for each probe in a panel of 2596 capture probes was calculated for all 16 normal cfDNA libraries that were profiled in this experiment. The average was then compared individually with three representative samples using scatter plot analysis. Each dot represents a different probe and its position on the graph is a comparison of the average on the x-axis and the individual sample on the y-axis. The tight diagonal distribution of the majority of probes reflected the highly-correlated unique read capture performance of most probes (R^2 correlation ≥ 0.95 for all three graphs). Importantly, the consistency of probe-by-probe sequencing depth supports the use of the targeted hybrid capture platform in copy number measurement.

[00246] With respect to copy number, the most straightforward treatment of probe data is to further normalize the adjusted genomic depth values that occur in autosomal chromosomes to a diploid-averaged value of “2”. The same is true for probe values that occur in females for X-linked loci. For X-linked and Y-linked regions in normal males, averaged copy values are appropriately set to “1”. This numerical transformation was applied to a set of chromosomal control probes (239 probes that target select loci on all 22 autosomal chromosomes, Table 3), a set of 199 probes that target the X-linked AR gene, and the 45 TP53-specific probes considered in detail above (FIG. 27A and 27B). Each dot represents the

value for an individual probe. With the exception of infrequent “noisy” probes, the vast majority of individual probe counts in regions anticipated to be diploid possessed values that were approximately “2”. Probes for the AR gene in a healthy male fluctuated with an average value close to the anticipated “1.”

Table 3: Chromosomal Control Probes

Name	Sequence	SEQ ID NO:
Chr_1_1	GTGTCTCGGCAACCACTCTTCACCAATATCACAGTGGACA	7734
Chr_1_2	ATCCAAGGGGAGGAGATCAGTGCCCCTATTTGTATCGCAC	7735
Chr_1_3	ACTTACTGAAGCAAGAACCTCATCAAGCTGCCTCCCACCA	7736
Chr_1_4	AGTTTGTGATCCTCTGTGGGCAACCTCAGCAGTCTGGTT	7737
Chr_1_5	GGAGAGCGGAGCTGCTCAGAGCTTGCCAGGTTCTAAGTG	7738
Chr_1_6	GACTGTGGCAATGAGGCAGCTAAGTGGTTCACCAACTTCT	7739
Chr_1_7	GGTGTATTTTGACAACGGTGGACCCAGACACTGGAGTCAT	7740
Chr_1_8	GTTGGTCTATTCTTGCGGTTGTAAAAGTGGCCCAGAGTGA	7741
Chr_1_9	GTGAGCCTTCTCTCACCATTCTGTCCAAAATAGCAGCCCT	7742
Chr_1_10	CAGCCTAGATATGATTCTCACTACCCTGTTCCATGGTTC	7743
Chr_1_11	AAAGAATGTGTTGGCTCATGATCAGACTTGAGCACTTGGG	7744
Chr_1_12	CCTAGGCTGTTGCTGCTGGACCTGTTTGTGCTTCATCACA	7745
Chr_2_1	CAGTTGACCCTTCAGCCACAGGGGTTTGAACCTTGAAGGA	7746
Chr_2_2	AGGACCTGAGTATGCACGTTTTTGGTATACTGGGTAGGGGT	7747
Chr_2_3	TATCAGCTGGGATGGTCCGGTCAGCAGCATTACCCTGTTT	7748
Chr_2_4	TGCCTGCTCAGCCCAGATTTTCAGTCATGCTGGCCATAAAC	7749
Chr_2_5	CTGGGGGGTGAGGTTTGAGGTTTGAGTGTGGGATGTGAGG	7750
Chr_2_6	CCAGCTTTTTTCAGAAGCTGGGAAAGTAATAACCCGTGTTG	7751
Chr_2_7	CCCAGCGCCCGTGGCTTTGGCTCCTCAGTCCCATTAAAT	7752
Chr_2_8	TATACCACCAAGTCTACCTACTGCCTGCACATGCTATGGC	7753
Chr_2_9	GGTCAATCCGGCACTACTGGTTGTCCAAAGGGAGGTTACT	7754
Chr_2_10	AATCAAACATCAGGACCGCCACAGCACAGGTCAATGAAC	7755
Chr_2_11	GTGTCTCCTGGAGGTGCATGGGTGGTTTTGAACTTCATTG	7756
Chr_2_12	GACCCATGTAAGGGGTTGGGTTATGTTCTCCTTTTGCCCA	7757
Chr_2_13	TCACTGACATGCGAAGCTGGGAACGAGAAAATGCACATCC	7758
Chr_2_14	TCCTACAGTGCTTAGGGATGAATCTGGCAAAGAAGGATGC	7759
Chr_2_15	GAAAGCAGTCCTTACCACAAGAAGACCCCGATGTGGTGGT	7760
Chr_2_16	ATTGCTCACTGGCTGGCTTGCATTTGGTATGCGATTGGGA	7761
Chr_2_17	GTCCCTGGGACCATCTGTGCATTGTTCTTGTAAGTGGAAA	7762
Chr_2_18	GACCGAATGGCGAACGCAGTGAATAGATCAGGAGGGAAAA	7763

Chr_3_1	GAAGGAATGGAGTGGAACAGATAGGGGTGAGGGAATAACG	7764
Chr_3_2	CCACTGCCATCCTCAGAGGGAGATTACAAAGTCTCACAAT	7765
Chr_3_3	ATCCAGGCTTCATGTTCAAATGCAATGGCCCTTGCCCCAT	7766
Chr_3_4	AAATTTCCCCTGGCTCCCTACTGCTTTGCAGGCCAAGTAA	7767
Chr_3_5	ACCTTAAAGACGGGCCCACATCTCTTTGGATGGGATTAGG	7768
Chr_3_6	GGGCTTCGGTTTTTGGCGAAGGTGCTCACAATCTTGATATC	7769
Chr_3_7	TGAGCTGTCCTTCATGCCTGCATTTCCCATGTCTGTCTTC	7770
Chr_3_8	ATCTTTATCCAGGGCTACCAGTGGTGGGTCCAAAATGACT	7771
Chr_3_9	TACAGGTGAAGGATGTCAACGAGTTTGCTCCCACCTTCAA	7772
Chr_3_10	GCTGTTGTGACGGAGGGCAAGATCTATGACAGCATTCTGC	7773
Chr_3_11	AATGAAGGGGATTCAAGCCTTGCCACCGACTTACAGGAAG	7774
Chr_3_12	TGTGAGCGTACTTTCTCCCCCAGGTTGAAGAGGAATGAGT	7775
Chr_4_1	ATTCCAAGTCCAGGTCCCAAATCTATCAGTACCGGCTGGC	7776
Chr_4_2	GACACAGAGTGCATGAAGACCGTTCAAATATGTCAGGGAC	7777
Chr_4_3	CATGAGTCCTTCTATGACTCCCTCTCAGACATGCAGGAAG	7778
Chr_4_4	TTTTTAGGAGACAGGTACCCACTGTCTGGTGACGAGGACT	7779
Chr_4_5	CCTTCTGTTGAGTCGCTAGGAGATGCCTCAGTTCAACAAT	7780
Chr_4_6	GACAGAACTTCATACCCAAGAGCTGCTTCTCAGCTGGA	7781
Chr_4_7	CAGGCAACTTTGGCAAGACCAAGTCAGCCTTCTCATCTCT	7782
Chr_4_8	CCCTTGCTACCATCACTGTTGTGTCATCTGTGCTTGCAATCC	7783
Chr_5_1	AGGTCTCACTCCAAGTGGCCCTGTATTAGAGCTAGGCTGC	7784
Chr_5_2	GAAACCATGCGGGATTCATCTTTGTGTCAGAGTGGAGCGGCA	7785
Chr_5_3	TATGAAATTAGGCGGTGGTTGGACGTGACTGTGTGTTGAC	7786
Chr_5_4	TGAAACTTGTCATGACATACTGCGGCTGCCCATTCACTAGG	7787
Chr_5_5	TGCTTCTTGTTTATAACTCCCCCTGGCCACCATCTCGGGCT	7788
Chr_5_6	ATTCCCTCTCATTTGTGGTTGGTGGCTGGATATCTGTTCC	7789
Chr_5_7	AGCATCAGCATTTCCCTGTGGACTTACCTCTCTCAGTAGT	7790
Chr_5_8	AAAATTTAAAGGTCGGCGGTAAGGCTGAAAGCCAACAGGC	7791
Chr_5_9	GAGTGTGTGCGGTCAGAAGGAACACCTGAGAAACCGCTTTA	7792
Chr_5_10	CATAGCAAATACCTGTCGCTGAGCCAGGAGTAAAGTCTGG	7793
Chr_5_11	AAGAGGCTCTGAGCTCTTGATAGAGGTTACATGGGGAGCA	7794
Chr_5_12	GGAGACAACCTTAGGAGGTTATCTAGACCATTCCCGCCTTC	7795
Chr_5_13	GTGTTTCCTCCCAGCATGCACTTTGTGGCTGCCTTTCTTT	7796
Chr_5_14	TGGCTTGTGTAGCGTGTTTCATTTTGGAACTTGGAGCCG	7797
Chr_5_15	GACACCTCTGGTGCAGTTTTGAGGCTGGCCGGAAGGGAT	7798
Chr_5_16	GTTTCAGATCTTGCAATGGGAGGGATCGACTCGGCCCTTT	7799
Chr_5_17	TGCCTAAATCAGAAATGGGCTACTTCCCTTGGCCACATCC	7800
Chr_5_18	CAATCTACCACCTCAAGGTTACGCGTGGAATTCTACACCT	7801

Chr_6_1	GAGTTTTTCTTTCAGGTAGTCTGAGATGGCCCCGCACCAAG	7802
Chr_6_2	TACTATAAAGAAGGCACCTCTAGGCTTGGCAAGCACACGT	7803
Chr_6_3	GGCAGATTCGATGGGACTTTAGACACTTGCTTTGCTCCCT	7804
Chr_6_4	CAAAATGTCCCCATGCAAACATGTCCCGCACTGTGTGGTAA	7805
Chr_6_5	ACATGTGTAATCTTCTTCTCCTAGGGCGGCAGAACTCATG	7806
Chr_6_6	CCCGAGGAAAGCTCCTCTTTGCTGACTGTAATGTACTGCA	7807
Chr_6_7	GAGGACAGCATTTCGCATATCAGGTCGAAATTTCTCCGCGA	7808
Chr_6_8	GTCCAGCTTTCATCCTTGATCCTGCTACTCTAGGCTCTCC	7809
Chr_6_9	ACTGATGGTGTTCACTTGCACCATCAGGTCTGATGGAGGA	7810
Chr_6_10	AATTGGTTCACAAAGCGTCGGGTGATCCAGTAACAGTCGA	7811
Chr_6_11	CAGAACTCTGCTCTAACGCCAAGCCTTCAATATGTCTTCG	7812
Chr_7_1	CAATTCTTACCATCCACAAATGGATCCAGACAACCTGTTT	7813
Chr_7_2	ACTACACCTCAGATATATTTCTTCATGAAGACCTCACAGT	7814
Chr_7_3	TGCTATAGACGCACAAACGACCGCGAGCCACAAATCAAGC	7815
Chr_7_4	CCATGACTTATGTGCAGCTTGCGCATCCAGGGGTAGATCT	7816
Chr_7_5	AGGAGTTGGTGGCTAAACCGCTGACTTTTCTATTGCAGAC	7817
Chr_7_6	GAAATATAACAGGACCAGAAGTGGCTCGCAGGAGACTCAT	7818
Chr_7_7	TAGCCAGACAGAAGGCGGACACTGATGATACCTCAAGACT	7819
Chr_7_8	GTTTGCCACCAGCGAAGAGAGCCATCCTGGTAGAATTGGA	7820
Chr_7_9	GGAGATATGCACTTGCCCTTTGGTAATCCTGCTCCTTCTG	7821
Chr_7_10	AAAATAACAGTAAGTACAGGGAGGGACCGAGAGGCATC	7822
Chr_7_11	AAGAACACCAGTCCATAAAGACGCATGTCCGGTGATGCCT	7823
Chr_7_12	AATCTGTTTAGACTGAGCAACTGTGCCAGCAGAGGGACCT	7824
Chr_8_1	AAGATGGCGAAGGTCTCAGAGCTTTACGATGTCACTTGGG	7825
Chr_8_2	CCATGCCTGCCAGCTGATAAGATTTGGTTACCTTTCCATG	7826
Chr_8_3	GCTGCAAGAAAGCGTAAGATTGCCATTCGAAAAGCCCAGG	7827
Chr_8_4	ATGCAGGAGTACAATGTGGGCATGTCCACCCTCTACGACA	7828
Chr_8_5	AGAACGGCTTTGCTGTCTTCCGGCAAACCTATGGTTCTGA	7829
Chr_8_6	TGGCTTTGGCGCTTTAAGGCCAGACACGGCATTAAAAAGC	7830
Chr_8_7	GCAGGCAGAGAAAGATGGCTTTAGAAACCTCTTCCCCACC	7831
Chr_8_8	TCAGCTGTGGCCATTGGTGGATCTCATCCTTAGTACTAGT	7832
Chr_8_9	CCATGGTTCTGTGAGACTGGTAGAAAGCACAGACCCCTTA	7833
Chr_9_1	AATGTGCTTATCACTCGTGATGGGGTCTGAAGCTGGCAG	7834
Chr_9_2	AGGGTCTCATTTTAAGACAGCTTGATTTGAGGGTGAGGGG	7835
Chr_9_3	CAGTTGCAAACCATACTTCCTTCAGCCCAGTCCTGTCTAT	7836
Chr_9_4	GTCTAAGGGCATCTTACCTCCAAGAAGTGTGAGGCGTA	7837
Chr_9_5	TACCTAGGGAATGACCACTAAGCACCATCTCCGTCACTCT	7838
Chr_9_6	GGAAGAGAGGAGGGTCATCCAGTCAGTTTTGCAGGAATCT	7839

Chr_9_7	TGCTGCAGTGTGCGGAAGAAACCTACCTGCGTTTCTTAGAA	7840
Chr_9_8	CATCATACCTATGGCATAGCCATCAGGGCACTGCAGTTTG	7841
Chr_9_9	TATATCTCACGTGACCGAGGATGGGTGCGTGGGCATTACACA	7842
Chr_9_10	GAAATGGCCATCTATAGGTGGGAACCACTCCAGTGTACACA	7843
Chr_10_1	GGAAACCTTTCAGTCTCTACTAGAAGCGCGGAGAGAACTC	7844
Chr_10_2	TCTGGCCGGCATTTCATTTAAGGCCTAAGGATGAAGGCGGT	7845
Chr_10_3	AGATACCCTATCGTTCCTTATCTCAGCGAAACAACCTCCCC	7846
Chr_10_4	CGCAACTCCTCCAGATCGCAGTGGTGCTTCTTCACTTTCA	7847
Chr_10_5	TGATTCCATGGTTGCCCCGTATACTCCATAAGGCGGTACTT	7848
Chr_10_6	ATACCATATCCGGCTTGTTAGGAGGAGGTATTACAGGGG	7849
Chr_10_7	GTACCTGTTAACCCAGACGCAATTCCTCCACAGTACACAG	7850
Chr_11_1	ATGTGACACTTGCATCCAGGGAGGTACCATCTGTGTATG	7851
Chr_11_2	CTAGGTCCTGAAGAGGTGGCAAGGAACCAGGACAGAACAT	7852
Chr_11_3	TCTGTCAATTGGTGACGCCATCTAGACTCTTGGCTTTGGGA	7853
Chr_11_4	AAGGTATAGAGCTGGGCGGCTTTCCTCGTTATAGGTGGAG	7854
Chr_11_5	CTCCTACGTAGCCGGGTAGAACTTATGGCAGAAGTCAGG	7855
Chr_11_6	TGGATTCCCAGGGTTAATTGTGACCCATTGCAGGAAGGTG	7856
Chr_11_7	AATGCTGTCCTACTATGGTCTGTACCTGTCCCAGAGGTGG	7857
Chr_11_8	GTGCACCTGGAGAGCATAACAGGGCACTGACTTGTAGATCA	7858
Chr_11_9	TTCCATCTCGCATAACCTGCCCCATAAACTCTTCTCGGTTC	7859
Chr_11_10	ATGAAGGCCTGCTTTGAGTTATCAGATAGGAAGGGGCCAG	7860
Chr_11_11	AGGTCATGTCCCGCTTTTGGCTGAACCTAGTTTTGCCCCAA	7861
Chr_12_1	CTGCATTCTCCATGAGTAGAGTACGAGCCTCATGTTGGTA	7862
Chr_12_2	AAGGCTGTCTTCACCAACTGGGTAGGTGTGGATCAAGACC	7863
Chr_12_3	CTGACTTTGGTGTTGGGGAGTCGGTGGTCTTCTTCCATT	7864
Chr_12_4	ACTGCAGAGGACCAGACTGGGAAAACAACGATATGGCAGG	7865
Chr_12_5	CCTGGCTTAGAAGTCTGGCCGGTCTTCTTCAGCTTCTTA	7866
Chr_12_6	AATCTCAGAAAGAGTTTCTGGGACCATGGCAAATGGTGCC	7867
Chr_12_7	ACATTATATCCGGTCCAGGAATATCTGGCTCAGGCTGGGT	7868
Chr_12_8	AAGCACAGGAAATGTGCCTCACACGACTTCACATGCCCTT	7869
Chr_12_9	GGGGGCTTTGCGGGAAGAGGGGACTAAACAACCCTTCTGT	7870
Chr_12_10	AAAAGAAATGCGATCAGCGCAACCCATCCGGTGTGGCGCT	7871
Chr_12_11	GGCAGTGGTACCATGACATACTTAGCAGAGATGGACTACA	7872
Chr_13_1	ATTTCCCATGCGAGAGGTAGCTTGCCCAGGCTGTTGGATA	7873
Chr_13_2	TTCCATGCCGAGTCCTGATGGAACTAGCACTGAAAGACC	7874
Chr_13_3	TCACGGGAGCTTCTTCACTGAGTTCTGCGAATCTGAAGC	7875
Chr_13_4	TTTCCAGAGATGAAGCACTACCCAGTCTTACCCAAGTTTCG	7876
Chr_13_5	CCACCGAGAACAGTGATGAAGGACTTAAAGTGAGAGATGG	7877

Chr_13_6	GTTCACTCGTCGGTTTTTCACCAACCACAGACTAGCCTCA	7878
Chr_13_7	ACGCAGCTGTGTTGAGTGCACAGGAAGCTCTTAGGGTTAA	7879
Chr_13_8	TCTCAGTGAACAGAGGGCTCACTGAGAGGACTTTGAATAC	7880
Chr_13_9	ATGGCACAGGCCACATACTGGAATGAATGACGGGCTTCAT	7881
Chr_13_10	TGCTGCTTGATGGTGGCATCACTGTCCCCTCATTCCATGA	7882
Chr_14_1	GGACACATGTGGACAGTGTGAAACCTCAGAACACTAACCC	7883
Chr_14_2	AAGTTCTTATCCTTAGGGACCCAGCGGAGACCTTGGTTCT	7884
Chr_14_3	CGACGATGCCTGGGAATAGGATCCATGGGATTGATGAGAA	7885
Chr_14_4	GGGAGCCATGAAGATTTCTCCCAGCTCCTGAGGAACTTTG	7886
Chr_14_5	TCTGGTCCCTCAAGTCCTCAGCTGTAGAAGTTCTCATTGCG	7887
Chr_14_6	TGCCAACCCCTGGAAACTGGCTTGTGTGTCCACAACAGAAA	7888
Chr_15_1	TAGGTGACAGCACTGTCCTTTCCTGCCATTTGCAGGGAA	7889
Chr_15_2	TTCTTCTAGATGGCAGACATTGTTGAGGCCTCCCGTACCT	7890
Chr_15_3	AGAGAGCTGCGAGACAAGACTTGAGTGCGACAAGATTTC	7891
Chr_15_4	TTCAATCAGGTACTCCGAGTTCCTTGGAGGCCAAAAGGA	7892
Chr_15_5	AGGAATATGGGGTCCATCTGAGACTCGCAAGTGATGATAC	7893
Chr_15_6	GATCTCCAGGACCAGCTCTCAGAAATGCACGATGAACTGG	7894
Chr_15_7	ACAGTGTGATGGAGCAGCAGTCCAAGTTCATCCTCCAAGA	7895
Chr_15_8	AAGATGACAGGATCCAGGAAACAAGACGCATGGGCCAGAA	7896
Chr_15_9	AAAGAGTGGGTCTGTTAATAATCAGGCCGAGACCACCAGC	7897
Chr_15_10	CACCCTTGTTTCGTGGCCCTTGCTTGGTAAACTGGTATCCA	7898
Chr_15_11	CCCAAGTATGGGTGAGGATGCTAGAAATGCCACATAATG	7899
Chr_15_12	AAGACTGTCATTGGTAGGTCATGATCCTTGGCAGCATGAC	7900
Chr_16_1	GTGGGGACGGTCATTATCAGCTTTCTGGACACACAGACAG	7901
Chr_16_2	TGAGAGGCCAAAGAATATCAGTTGACTCTGGATCAGGGGC	7902
Chr_16_3	GAGGCTTTTTAGGGCAGCGAGAAAACGGGAACCTTCATTCC	7903
Chr_16_4	AGGACTTCTCTGGACCTGTGCCTCACTACTCACCTGGAT	7904
Chr_16_5	TGGCCACAAATGTTGCCTCCAGCTGCTCAATGTTCTCCAA	7905
Chr_16_6	CTGGCATTGGTGAGTAATAGGAGCCAGACGGGTCTGTGTT	7906
Chr_16_7	ATACTTACCTGCACGAGAATGAGTTTGGAGCGCAAGGGGG	7907
Chr_16_8	TTCCCCCAGAGACTCTGTCCACTATGGACATTAAAAATGTG	7908
Chr_16_9	GTGCTACCCTCCTCCCTCAGGTTATGTGGTCCAGGCTTT	7909
Chr_16_10	TAAGTGGAACAACATTCCCTTCATTATAGCCCTTCGTGGG	7910
Chr_16_11	GCAACGTCAACAATACTACTACGTGCACAAGCGCCTCTACTG	7911
Chr_17_1	GCGGATGTCGTTATGGGACAGGTACAAGTAGATAAGTTGC	7912
Chr_17_2	GTGGTCACCATCTCTTCAAACCATTTGGACTGGGCCTGGT	7913
Chr_17_3	AAGCCAAGGAGTTCTGAGAGAGCTTAGCTAAGTTCTTCGC	7914
Chr_17_4	TTTTTTAGTACCCCAGTGTGTAAGACCAACTGAGGGTGGC	7915

Chr_17_5	GTTGTCATTGGGGCTATAGACATAAGCACCTTCCGGAATC	7916
Chr_17_6	CTGAGTGTGCGAGGGGAAGATATTGGTGAAGACCTGTTCT	7917
Chr_17_7	GTCAGACCCTGTCCTCGTCTCCTTTACCTTGTCTCGATTT	7918
Chr_17_8	TAAACTATGCTCGCCACCACTCAGCACTCACCTCTTGGGC	7919
Chr_17_9	GGCAACTTCCTGAGACAGATCGGTAAAAACAACCCCTTCT	7920
Chr_17_10	TCAACTGTATTTTCATCAGAGAGATGTGGCTTTCCAGACA	7921
Chr_17_11	GTTTCCCTCATGTTCCCCCAGGTTCTGTCAGGTGAAGCTG	7922
Chr_18_1	TTAACCCATCTCTACCCGTCTGTGTCAAGAACGGAGGCT	7923
Chr_18_2	CTGCCCCAAAATAGAAACCGAGGTTCTCCGTGACCTACATC	7924
Chr_18_3	TTCTTTGTCAGTAACAGCGGGAACATGAAGCCGCCACTCT	7925
Chr_18_4	TGGTTTGCCAGTTCAGACACCCAGCCAAATTGCCCTCTCA	7926
Chr_18_5	TAGTGCAGCTGGCTTTGAGCCTGTTCCCGAATGTTTCAGAT	7927
Chr_18_6	AGGGTAATAGCACCAAGCTCTAGTCTACCCACCTCTCTGA	7928
Chr_18_7	CCGCATCTCTGGAGTAGGAATTGATCAGCCACCATATGGG	7929
Chr_18_8	CTATGAGCATACTGGGGAGGGAAACCTCTAAGCGGAACTT	7930
Chr_18_9	AAAAACCTGCAGGAAGGAGACCTGAATGCAACTGTGGGTC	7931
Chr_18_10	CAGGTGCTCCAAACCTTCCAGTCTATGTTGTAGATTGCAG	7932
Chr_18_11	GCCATACTAACCTACTTCTCCTTGAAGCTCTTGGCCCATC	7933
Chr_19_1	ACTGTGAGATAGCCCTCATCATCTTCAACAGCGCCAACCG	7934
Chr_19_2	AGATACACGGTCACAGACGCCATGTGTTGTGGCTTCTGCA	7935
Chr_19_3	CACATCCTCTCACCTTTTCCGAAGGTTGCAGCTCCTTCTC	7936
Chr_19_4	TCTGTCTCACCGGTCCCTTCATTCTAGGCAACTGTAGAT	7937
Chr_19_5	ATATCATGGTCTGTATCCCCCAGGTACCTTGACACAGGCC	7938
Chr_19_6	CTCTCCGCCTTTCTTTAGACCTGAGCATGCAGAATTCCGA	7939
Chr_19_7	AAGGCATTTAAATGGGACAGCGTCCCATGCGTGACTTCTC	7940
Chr_19_8	TCTTTCTAACAGACGAACAGCCTACACCTACAACCCCGAG	7941
Chr_19_9	GTCCAGCCCCAAAAGCATCTTGGGTAAGGATTTGGGATCA	7942
Chr_19_10	GTTGTTCTGGGCCAGTGTTAGTTGCTCACATGTCCTGTCT	7943
Chr_19_11	AACATGCCTCTTAGTCCTGGGCCATACCTTAGCCTTGTGC	7944
Chr_20_1	TAACCTCCAAAAGAGGTACCCATTGGCGCTCAACCGAATT	7945
Chr_20_2	CTATATCTCCGACTATGCCTTCTTGGGCACTGCACTGCTG	7946
Chr_20_3	TCTAGATGGAAGCTGTATCCAAGGATGCTCCGGAATGTTG	7947
Chr_20_4	ATCTTCTCTGCCTGCCGCACTAGCTTCTTGGTGACTTCTC	7948
Chr_20_5	ATCGAGTTGTGCGAGCCCCATGATTCGACACCAAGATCCCA	7949
Chr_20_6	AGGTGCTTGTGTTTACTCTCTCCAGGTGATGATGCCAGGGA	7950
Chr_20_7	GTGCACTGTCAGATCTTGAAACGGCCAAAGGATTTTTCC	7951
Chr_20_8	CATTTTGCAGGAGGCTGCTAATTAAGGCTGAGGGCCATCA	7952
Chr_20_9	TCAATGGTAGACTGGAGTACCTTGCCAGGGCAGAGAAAAA	7953

Chr_20_10	CTCCTCCAGGAGCTGGCAGCATCAAGACCCCACTTCGCTT	7954
Chr_21_1	AAATAATAGCAGGCGTTGAGATGTCCCTTCCCCAGCACTC	7955
Chr_21_2	AAGTCTGACAGCATCTGCTTGAAGTGAAGCACAGTGATGG	7956
Chr_21_3	ATTCGTGATGGCGCTCATTTCATAAAGGACGACAGGTCA	7957
Chr_21_4	GAAGAGTGAATTCCCGCTTCTGCGCCAACATTCTGTTTCC	7958
Chr_21_5	ACAGGTGAAGTCTTTGCGTGCCTCCCTGTTGGACTCAAAT	7959
Chr_21_6	TAATGATATTCTGGCACAAGGAGCAGAGCCCCTCTTCTTC	7960
Chr_21_7	AGACCCAGCCTACCTGCATGATCTCTTGTACAGCTTTGCA	7961
Chr_21_8	TCATGGAACATGGGCCTTGCAAAGGGGTCAAGATCACAAAC	7962
Chr_21_9	GTCAAAAAGGTCCAATCAGCTAGAGACTAGGCCAGACCCA	7963
Chr_22_1	TGTGACCACCCTAAAGGGAGGGCAGAAGCCGAGTCACCCT	7964
Chr_22_2	ACGCCTCCACCTGCTGCTAGGACTCCCCTCCCAAACAAAG	7965
Chr_22_3	CACAGTCTAGACCCTGATGGGCGATCTCAGTAGTGCTGTT	7966
Chr_22_4	CCTATCAACGTGCAAGTGCGATTGTCTCCACTGGCTTTC	7967
Chr_22_5	GAAAATCATTCCCCATTCTGCAGGATCCGTTCCCCTGGCA	7968
Chr_22_6	AGTGGGACATACCAACTTGATGAGGCAGTTGTGCGAGTTC	7969
Chr_22_7	GTAAACAGCTGTCTTCTTACCCTACAGATCATTGGGCAGG	7970
Chr_22_8	CAGAAGGATACTAGAATGGAATGTCCTGCGTGACGAAAGC	7971
Chr_22_9	AGTTCACATCTGATTCTCCTATGGCTGCTAGGCTCCAGGA	7972

[00247] Significantly, when the same analysis was applied to cfDNA collected from the blood plasma fraction of a castration-resistant prostate cancer patient using healthy samples as normalization controls, three prominent features emerged (FIG. 27C). First, all of the control probes exhibited noisy counting behavior. Second, the counts across all AR probes were significantly elevated from a normal value of “1” to an amplified value of approximately “5”. Amplification of the AR gene is consistently observed in advanced prostate cancer patients. Third, the TP53 probe counts, while more tightly clustered, possessed an average value far closer to “1” than the expected value of “2.” This likely reflected inactivation of one or both alleles of TP53 by copy number loss in the fraction of circulating DNA derived from tumor tissue.

[00248] These data indicated that the methods of the present invention comprise three important karyotyping aspects. Namely, the methods described herein detect generalized chromosomal aneuploidy, copy increases of specific, targeted genes, and copy losses in the same specific, targeted genes. These result further indicate that the methods and

platforms described herein can guide the use of precision therapies, as all three of these genomic abnormalities occur frequently in cancer.

[00249] Generalized chromosomal aneuploidy for castration-resistant prostate cancer patient samples (blue dots) relative to a healthy control (brown dots) was measured (FIG. 28). In this analysis, the approximate ploidy for all 239 control probes used in the experiment were ordered according to their chromosomal targets. For some chromosomes (*e.g.*, chromosome 1 and chromosome 22) a similar ploidy value of “2” was observed between patient and control samples. In other cases, deviation between the two samples was observed. The degree of information regarding overall genomic ploidy provided by these experiments was constrained by the number and density of control probes used. However, these data indicate that a denser probe panel covering all chromosomal segments at uniform density can be used – in conjunction with the additional unique features of the present invention. Such analyses will provide a higher resolution, genome-wide measurement of chromosomal copy number.

[00250] These data further highlight the capabilities of the present invention as a guide for precision therapy. For example, tumors that possess genomic deficiencies in homologous recombination repair often exhibit highly destabilized chromosomal ploidies, and patients with such tumors are good candidates for inhibitors of the PARP enzyme complex (See Popova *et al.*, Genome Biol. 2009;10(11):R128). Unlike most sequencing assays that seek to *genotype* a tumor, the assays described herein use sequencing to detect destabilized chromosomal ploidy as a *tumor phenotype*, even if the causal mutations driving this phenotype remain hidden from targeted analysis.

[00251] The ability to detect gene loss in DNA shed from solid tumors is especially significant. Mutation and deletion of tumor suppressor genes is a frequent event in cancer genomes; moreover, individuals with germline loss of tumor suppressor genes are uniquely vulnerable to developing cancer later in life. The diagnostic value of a liquid biopsy copy number loss (CNL) assay is directly proportional to its sensitivity. To determine the lower limit of detection for the invention described here, the immortalized lines described in Example 1 were systematically diluted into the “genome-in-a-bottle” reference cell line, NA12878. One line had a single copy deletion (monoallelic loss) of ATM, the other a single copy deletion of BRCA2. The experiment included four control samples of pure NA12878 and eight spike-in samples containing 16% of each monoallelic deletion line (FIG. 29). For reporting purposes, this corresponds to an 8% minor allele frequency of biallelic loss.

Averaged values for all probes targeting specific genes and two additional, undeleted control genes are shown in FIG. 29. Copy loss of ATM and BRCA2 was confined to spike-in samples only. Additional computational treatment of the data revealed confident copy loss calling of biallelic deletions down to 2% minor allele frequencies. This sensitivity indicated that the present invention required no specialized considerations in order to routinely include copy loss calls in standard blood-based genotyping assays.

[00252] These data demonstrate the use of probe-specific genomic capture data for the analysis of copy number, including both copy number gain and copy number loss of target genomic loci. Additionally, the invention described herein has been shown to possess the sensitive ability to detect single nucleotide variants, insertions and deletions ranging from single nucleotides to many thousands of base pairs, and gene fusions resulting from chromosomal rearrangement by aberrant mutational processes (See PCT Publication No. WO 2016/028316; and U.S. Patent Publication No. 2014-0274731). All of these mutational processes can contribute to the transformation of normal tissue to neoplastic cancers, and as precision therapies continue to emerge, accurate diagnosis of these diseased genomic signatures will become an increasingly indispensable feature of precision medicine.

CLAIMS

1. A method for performing a genetic analysis on a DNA target region from a test sample comprising a plurality of genomic DNA fragments, the method comprising:

(a) generating a genomic DNA library by contacting the test sample with a set of adaptors, wherein each adaptor of the set of adaptors comprises a sample tag region selected from a pool of unique sample tag regions, wherein the pool is selected from a plurality of pools, and wherein the selected pool is unique to the test sample; wherein the genomic DNA library comprises a plurality of library DNA fragments and at least two unique sample tag regions; wherein each of the library DNA fragments comprises a genomic DNA fragment from the test sample attached to an adaptor;

(b) contacting the genomic DNA library with a plurality of capture probes that specifically bind to the DNA target region, thereby forming complexes between the capture probes and library DNA fragments comprising the DNA target region; and

(c) performing a quantitative genetic analysis of the genomic DNA fragments comprising the DNA target region;

wherein the adaptor is a DNA polynucleotide that comprises: (i) an amplification region, (ii) a sample tag region, and (iii) an anchor region;

wherein the amplification region comprises a polynucleotide sequence capable of serving as a primer recognition site for PCR amplification; wherein the sample tag region identifies the test sample;

and wherein the anchor region comprises a polynucleotide sequence that is capable of attaching to the genomic DNA fragment.

2. The method of claim 1, wherein the method identifies a genetic change indicative of a disease state selected from a single nucleotide variant (SNV), an insertion less than 40 nucleotides in length, a deletion of a DNA region less than 40 nucleotides in length, and/or a change in copy number.

3. The method of claim 2, wherein the genetic change indicative of a disease state is a change in copy number.

4. The method of claim 1, wherein the test sample is a tissue biopsy.

5. The method of claim 4, wherein the tissue biopsy is taken from a tumor or a tissue suspected of being a tumor.
6. The method of claim 1, wherein the genomic DNA fragments are cell free DNA (cfDNA) or cellular DNA.
7. The method of claim 6, wherein the genomic DNA fragments are isolated from the test sample; and wherein the test sample is a biological sample selected from the group consisting of: amniotic fluid, blood, plasma, serum, semen, lymphatic fluid, cerebral spinal fluid, ocular fluid, urine, saliva, stool, mucous, and sweat.
8. The method of claim 1, wherein the genomic DNA fragments are obtained by the steps comprising;
 - (i) isolating cellular DNA from the test sample;
 - (ii) fragmenting the cellular DNA to obtain the genomic DNA fragments.
9. The method of claim 8, wherein step (ii) is performed by contacting the cellular DNA with at least one digestion enzyme.
10. The method of claim 8, wherein step (ii) is performed by applying mechanical stress to the cellular DNA.
11. The method of claim 10, where the mechanical stress is applied by sonicating the cellular DNA.
12. The method of claim 1, wherein the amplification region is between 10 and 50 nucleotides in length.
13. The method of claim 1, wherein the amplification region is between 20 and 30 nucleotides in length.
14. The method of claim 1, wherein the amplification region is 25 nucleotides in length.

15. The method of claim 1, wherein the sample tag region is between 5 and 50 nucleotides in length.
16. The method of claim 15, wherein the sample tag region is between 5 and 15 nucleotides in length.
17. The method of claim 15, wherein the sample tag region is 8 nucleotides in length.
18. The method of claim 1, wherein the adaptor further comprises a unique molecule identifier multiplier (UMI multiplier) that is adjacent to or contained within the sample tag region.
19. The method of claim 18, wherein the UMI multiplier is between 1 and 5 nucleotides in length.
20. The method of claim 18, wherein the UMI multiplier is 3 nucleotides in length, and comprises one of 64 possible nucleotide sequences.
21. The method of claim 1, wherein the anchor region is between 1 and 50 nucleotides in length.
22. The method of claim 21, wherein the anchor region is between 5 and 25 nucleotides in length.
23. The method of claim 21, wherein the anchor region is 10 nucleotides in length.
24. The method of claim 1, wherein step (a) comprises attaching a first adaptor to one end of a genomic DNA fragment of the test sample and attaching a second adaptor to the other end of the genomic DNA fragment.
25. The method of claim 24, wherein the genomic DNA fragment is end repaired prior to attaching the first and second adaptors.

26. The method of claim 1, wherein the amplification region of each adaptor of the set of adaptors is identical to the amplification region of every other adaptor of the set of adaptors.

27. The method of claim 1, wherein the pool of sample tag regions comprises between 2 and 1,000 unique sample tag region sequences.

28. The method of claim 27, wherein the pool of sample tag regions comprises between 50 and 500 unique sample tag region sequences.

29. The method of claim 27, wherein the pool of sample tag regions comprises between 100 and 400 unique sample tag region sequences.

30. The method of claim 27, wherein the pool of sample tag regions comprises between 200 and 300 unique sample tag region sequences.

31. The method of claim 27, wherein each sample tag region of the pool of sample tag regions is 8 nucleotides in length.

32. The method of claim 27, wherein the pool of sample tag regions comprises 240 unique sample tag region sequences.

33. The method of claim 27, wherein each unique sample tag region sequence is discrete from any other unique sample tag region sequence by Hamming distance of at least two.

34. The method of claim 1, wherein the sample tag region identifies the genomic DNA fragment attached thereto.

35. The method of claim 34, wherein the adaptor further comprises a UMI multiplier which increases the number of unique sequences for identifying the genomic fragments.

36. The method of claim 1, wherein the anchor region of each adaptor of the set of adaptors comprises one of four sequences, and wherein each sample tag region of a given sequence is paired to only one of the four anchor regions of a given sequence.

37. The method of claim 1, wherein the amplification region of each adaptor of the set of adaptors comprises an identical nucleotide sequence;

wherein each sample tag region of the pool of sample tag regions is 8 nucleotides in length, wherein each unique sample tag region sequence is discrete from any other unique sample tag region sequence by Hamming distance of at least two,

wherein each adaptor of the set of adaptors comprises a UMI multiplier that is adjacent to or contained within the sample tag region, wherein the UMI multiplier of each adaptor of the set of adaptors is three nucleotides in length, and wherein the UMI multiplier of a given sequence is paired to one sample tag region of a given sequence,

wherein the anchor region of each adaptor of the set of adaptors comprises one of four nucleotide sequences, and wherein each sample tag region of a given sequence is paired to only one of the four anchor regions of a given sequence.

38. The method of claim 24, wherein the step of attaching each of the first and second adaptors to the genomic DNA fragment comprises:

(i) attaching an oligonucleotide comprising at least a portion of an anchor region to the genomic DNA fragment, wherein the oligonucleotide is part of a DNA duplex comprising a 5' phosphorylated attachment strand duplexed with a partner strand, wherein the partner strand is blocked from attachment to the genomic DNA fragment by chemical modification at its 3' end, and wherein the attachment strand is attached to the genomic DNA fragment; and

(ii) contacting the genomic DNA fragment attached to the attachment strand with a DNA oligonucleotide encoding a full-length adaptor sequence, a T4 polynucleotide kinase, a Taq DNA ligase, and a full-length Bst polymerase under conditions suitable for DNA ligation;

thereby attaching each of the first and second adaptors to the genomic DNA fragment.

39. The method of claim 1, wherein the DNA target region is analyzed for a change in copy number.

40. The method of claim 1, wherein step (c) comprises purification of the complexes formed between the capture probes and DNA library fragments comprising the DNA target region.

41. The method of claim 1, wherein step (c) comprises purification of the complexes formed between the capture probes and DNA library fragments comprising the DNA target region and performing primer extension and/or amplification of the DNA library fragments comprising the DNA target region.

42. The method of claim 1, wherein step (c) comprises purification of the complexes formed between the capture probes and DNA library fragments comprising the DNA target region and performing primer extension and amplification of the DNA library fragments comprising the DNA target region.

43. The method of claim 1, wherein step (c) comprises DNA sequencing of the DNA library fragments comprising the DNA target region to generate a plurality of sequencing reads.

44. The method of claim 1, wherein the genomic analysis comprises determining a change of a copy number in a DNA region of interest, and wherein step (c) comprises:

(i) determining a copy number of the region of interest present in the genomic DNA library derived from the test sample, and

(ii) comparing the copy number determined in step (i) to a copy number of the region of interest present in a genomic DNA library derived from a reference sample, wherein the reference sample comprises a known copy number of the DNA target region.

45. The method of claim 44, wherein determining the copy number in the region of interest comprises DNA sequencing of the DNA library fragments comprising the DNA target region to generate a plurality of sequencing reads, wherein each sequencing read comprises a unique molecular identification element (UMIE).

46. The method of claim 45, wherein the UMIE comprises sequencing information from the adaptor and at least a portion of the genomic DNA fragment sequence.

47. The method of claim 46, wherein sequencing reads comprising identical UMIEs are identified as a unique genomic sequence (UGS).

48. The method of claim 44, further comprising determining a raw genomic depth (RGD) for each of the capture probes contacted with the genomic DNA library.

49. The method of claim 48, wherein determining the RGD comprises determining the average number of UGSs associated with each capture probe sequence within a group of sample replicates.

50. The method of claim 49, wherein capture probes associated with a highly variable number of UGSs are identified as noisy probes and are removed from further calculations.

51. The method of claim 49, further comprising calculating an RGD for a sample, comprising calculating a numerical average of all RGDs for all capture probes in the sample.

52. The method of claim 49, wherein the RGD values for noisy probes are not included in calculating an RGD for a sample.

53. The method of claim 48, wherein the RGDs for the capture probes are normalized across all samples in an experimental group by converting the RGD for each capture probe into a probe-specific, normalized read count comprising:

- (i) multiplying each capture probe RGD in a sample by a normalization constant, wherein the normalization constant comprises any real number; and
- (ii) dividing the product of (i) by the RGD calculated for the corresponding sample; or
- (iii) dividing the product of (i) by an average RGD calculated from a subset of probes.

54. The method of claim 53, wherein the subset of probes is a set of control probes.

55. The method of claim 54, wherein the probe-specific, normalized read counts are converted to a copy number value comprising:

(i) multiplying the probe-specific, normalized read counts of probes directed to autosomal and/or X-linked regions by 2 in samples derived from females;

(ii) multiplying the probe-specific, normalized read counts of probes directed to Y-linked and/or X-linked regions by 1 in samples derived from males;

(iii) averaging the products of (i) and/or (ii) across all samples in an experiment; and

(iv) dividing the product of (i) and/or (ii) by the average of (iii).

56. The method of claim 55, wherein the approximate copy number values for all probes that target a specific gene are averaged.