(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
22 June 2017 (22.06.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/106768 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01) *G06F 19/10* (2011.01)
C12M 1/34 (2006.01) *G01N 33/48* (2006.01)

(21) International Application Number:
PCT/US2016/067356

(22) International Filing Date:
16 December 2016 (16.12.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/269,051 17 December 2015 (17.12.2015) US

(71) Applicant: **GUARDANT HEALTH, INC.** [US/US]; 505 Penobscot Dr., Redwood City, California 94063 (US).

(72) Inventors: **ELTOUKHY, Helmy**; 2 Barry Lane, Atherton, California 94027 (US). **TALASAZ, AmirAli**; 2181 Camino a Los Cerros, Menlo Park, California 94025 (US). **CHUDOVA, Darya**; 5931 Taormino Avenue, San Jose, California 95123 (US). **ABDUEVA, Diana**; 227 Orchard Road, Orinda, California 94563 (US).

(74) Agent: **AMODEO, Gabriele A.**; Wilson Sonsini Goodrich & Rosati, 650 Page Mill Road, Palo Alto, California 94304 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

— *with international search report (Art. 21(3))*

(54) Title: METHODS TO DETERMINE TUMOR GENE COPY NUMBER BY ANALYSIS OF CELL-FREE DNA

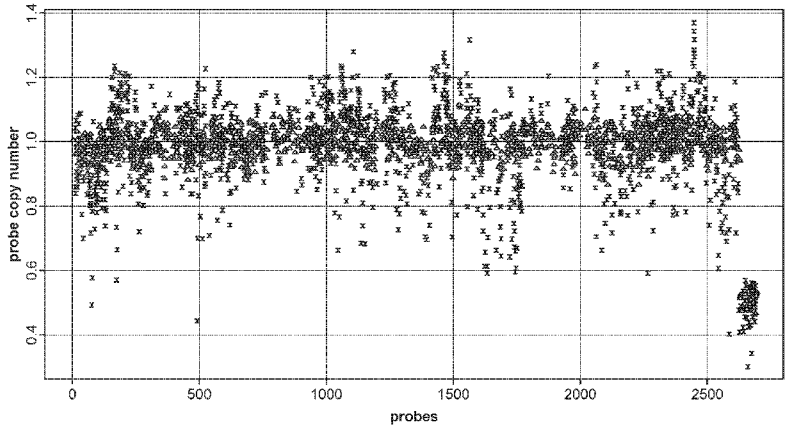


FIG. 15

(57) Abstract: Methods are provided herein to improve automatic detection of copy number variation in nucleic acid samples. These methods provide improved approaches for determining baseline copy number of genetic loci within a sample, reduce variation due to features of genetic loci, sample preparation, and probe exhaustion.

WO 2017/106768 A1

METHODS TO DETERMINE TUMOR GENE COPY NUMBER BY ANALYSIS OF CELL-FREE DNA

CROSS-REFERENCE

[0001] This application claims priority to U.S. Provisional Application No. 62/269,051, filed December 17, 2015, which is hereby incorporated by reference in its entirety.

BACKGROUND

[0002] Cancer is caused by the accumulation of mutations within an individual's normal cells, at least some of which result in improperly regulated cell division. Such mutations commonly include copy number variations, in which the number of copies of a gene within a tumor genome increases or decreases relative to the subject's noncancerous cells.

[0003] Detecting and characterizing copy number variation in tumor cells is used to monitor tumor progression, predict patient outcome, and refine treatment choices. Conventional methods, however, are performed on cellular samples that are often obtained by painful and time-intensive biopsies. Such biopsies also can often only examine a fraction of the tumor cells within a subject, and thus are not always representative of the population of tumor cells. There is a need for simpler, more rapid tests for copy number variation in tumors that do not require cellular biopsies, fluorescent in situ hybridization (FISH), comparative genome hybridization arrays, or quantitative fluorescent polymerase chain reaction (PCR) assays.

[0004] A particular challenge in determining copy number variation using sequencing data is that genetic loci will exhibit variance in their depth of coverage for reasons unrelated to true copy number. For example, amplification efficiency, PCR efficiency, and guanine-cytosine content can cause differing depths of coverage even for individual genetic loci present in the sample at the same copy number. Improved methods of removing bias due to such effects are needed to improve copy number detection.

SUMMARY

[0005] There exists a considerable need for improved methods to detect copy number variation in tumor cells from samples derived from cell-free bodily fluids. The present invention addresses this need and provides additional advantages. In one aspect, the present disclosure provides a method comprising: (a) obtaining sequencing reads of deoxyribonucleic acid (DNA) molecules of a cell-free bodily fluid sample of a subject; (b) generating from the sequence reads a first data set comprising for each genetic locus in a plurality of genetic loci a quantitative measure related to sequencing read coverage ("read coverage"); (c) correcting the first data set

by performing saturation equilibrium correction and probe efficiency correction; (d) determining a baseline read coverage for the first data set, wherein the baseline read coverage relates to saturation equilibrium and probe efficiency; and (e) determining a copy number state for each genetic locus in the plurality of genetic loci relative to the baseline read coverage. In some embodiments, the first data set comprises, for each genetic locus in a plurality of genetic loci, a quantitative measure related to (i) guanine-cytosine content ("GC content") of the genetic locus. In some embodiments, the method comprises, prior to (c), removing from the first data set genetic loci that are high-variance genetic loci, wherein removing comprises: (i) fitting a model relating the quantitative measures related to guanine-cytosine content and the quantitative measures of sequencing read coverage of the genetic loci; and (ii) removing from the genetic loci at least 10% of the genetic loci, wherein the removing the genetic loci comprises removing genetic loci that most differ from the model, thereby providing the first data set of baselining genetic loci. In some embodiments, the method comprises removing at least 45% of the genetic loci.

[0006] In some embodiments, performing saturation equilibrium correction comprises transforming the first data set of baselining data genetic loci into a saturation corrected data set by: (i) determining for each genetic locus from the first data set of baselining genetic loci a quantitative measure related to the probability that a strand of DNA molecule from the sample derived from the genetic locus is represented within the sequencing reads; (ii) determining a first transformation for the read coverage by relating the read coverage of the first data set of baselining genetic loci to both the GC content of the first data set of baselining genetic loci and the quantitative measure related to the probability that a strand of DNA derived from each locus in the first data set of baselining genetic loci is represented within the sequencing reads; and (iii) applying the first transformation to the read coverage of each genetic locus from the first data set of baselining genetic loci to provide the saturation corrected data set, wherein the saturation corrected data set comprises a first set of transformed read coverages of the first data set of baselining genetic loci.

[0007] In some embodiments, determining the first transformation comprises (i) determining a measure related to central tendency of the read coverage of the first data set of baselining genetic loci; (ii) determining a function that fits the measure related to central tendency of the read coverage of the first data set of baselining genetic loci based on the GC content of the genetic locus and the quantitative measure related to the probability that a strand of DNA derived from the genetic locus is represented within the sequencing reads; and (iii) for each genetic locus of the first data set of baselining genetic loci, determining a difference between the read coverage predicted by the function and the read coverage, wherein the difference is the

transformed read coverage. In some embodiments, the function is a surface approximation. In some embodiments provided herein, the surface approximation is a two-dimensional second degree polynomial.

[0008] In some embodiments, performing probe efficiency correction comprises transforming the saturation corrected data set into a probe efficiency corrected data set by: (i) removing from the saturation corrected data set genetic loci that are high-variance genetic loci with respect to the first set of transformed read coverages, thereby providing a second data set of baselining genetic loci; (ii) determining a second transformation for the first set of transformed read coverages related to the probe efficiency of the second data set of baselining genetic loci; and (iii) transforming the first set of transformed read coverages of the second data set of baselining genetic loci with the second transformation, thereby providing the probe efficiency corrected data set, wherein the probe efficiency corrected data set comprises a second set of transformed read coverages of the second data set of baselining genetic loci. In some embodiments, removing from the first data set genetic loci that are high-variance genetic loci comprises: (i) fitting a model relating the GC content and the first set of transformed read coverages of the saturation corrected data set; and (ii) removing from saturation corrected data set at least 10% of the genetic loci, wherein the removing the genetic loci comprises removing genetic loci that most differ from the model, thereby providing the second data set of baselining genetic loci. In some embodiments provided herein, the removing is at least 45% of the genetic loci.

[0009] In some embodiments, probe efficiency is determined by performing the saturation equilibrium correction on one or more reference samples, wherein the probe efficiency is the transformed read coverage obtained by performing the saturation equilibrium correction. In some embodiments, one or more reference samples are cell-free bodily fluid samples from a subject without cancer. In some embodiments provided herein the one or more reference samples are cell-free bodily fluid samples from a subject with cancer, wherein the corresponding genetic locus has not undergone copy number alteration.

[0010] In some embodiments, determining the second transformation comprises (i) fitting the probe efficiency determined for the genetic loci from the one or more reference samples to the first set of read coverages from the second data set of baselining genetic loci; (ii) dividing the transformed read coverages of each genetic locus of the second data set of baselining genetic loci by a predicted probe efficiency based on the fitting of (i). In some embodiments, the method further comprises: (f) determining a third transformation for the second set of transformed read coverages by relating the transformed read coverages of the second data set of baselining genetic loci to both the GC content of the second data set of baselining genetic loci and the quantitative measure related to the probability that a strand of DNA derived from the

each locus in the second data set of baselining genetic loci is represented within the sequencing reads; and (g) applying the third transformation to the second set of transformed read coverages to provide a fourth data set, wherein the fourth data set comprises a third set of transformed quantitative read coverages.

[0011] In some embodiments, the DNA of the cell-free bodily fluid sample is enriched for the set of genetic loci using one or more oligonucleotide probes that are complementary to at least a portion of the genetic locus from the set of genetic loci. In some embodiments, the GC content of each genetic locus from the set of genetic loci is a measure related to central tendency of guanine-cytosine content of the one or more oligonucleotide probes that are complementary to at least a portion of the genetic locus from the set of genetic loci. In some embodiments, the read coverage of the genetic locus is a measure related to central tendency of the read coverage of regions of the genetic locus corresponding to the one or more oligonucleotide probes. In some embodiments, the performing saturation equilibrium correction and the performing probe efficiency correction comprises fitting a Langmuir model, wherein the Langmuir model comprises probe efficiency (K) and saturation equilibrium constant (Isat). In some embodiments, K and Isat are determined empirically for each oligonucleotide probe in the one or more oligonucleotide probes. In some embodiments, the performing saturation equilibrium correction and performing probe correction comprises fitting the read coverages of the genetic loci to the Langmuir model assuming that the genetic loci are present in identical copy number states, thereby providing a baseline read coverage. In some embodiments, the identical copy number states are diploid. In some embodiments the baseline rad coverage is a function dependent on the probe efficiency and the saturation equilibrium.

[0012] In some embodiments, determining a copy number state comprises comparing the read coverage of the genetic loci to the baseline read coverage. In some embodiments, the cell-free bodily fluid is selected from the group consisting of serum, plasma, urine, and cerebrospinal fluid. In some embodiments, the read coverage is determined by mapping the sequencing reads to a reference genome. In some embodiments, obtaining the sequencing reads comprises ligating adaptors to the DNA molecules from the cell-free bodily fluid from the subject. In some embodiments, the DNA molecules are duplex DNA molecules and the adaptors are ligated to the duplex DNA molecules such that each adaptor differently tags complementary strands of the DNA molecule to provide tagged strands. In some embodiments, determining the quantitative measure related to the probability that a strand of DNA derived from the genetic locus is represented within the sequencing reads comprises sorting sequencing reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-

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

[0013] In some embodiments, determining the read coverage comprises collapsing the sequencing reads based on position of the mapping of the sequencing reads to the reference genome and the barcode sequences. In some embodiments, the genetic loci comprise one or more oncogenes. In some embodiments, a method comprises determining that at least a subset of the baselining genetic loci has undergone copy number alteration in the tumor cells of the subject by determining relative quantities of variants within the baselining genetic loci for which the germline genome of the subject is heterozygous. In some embodiments, the relative quantities of the variants are not approximately equal. In some embodiments, baselining genetic loci for which the relative quantities of the variants are not approximately equal are removed from the baselining genetic loci, thereby providing allelic-frequency corrected baselining genetic loci. In some embodiments, the allelic-frequency corrected baselining genetic loci are used as the baselining loci in the methods of any one of the preceding claims.

[0014] In another aspect, the present disclosure provides a method comprising: receiving into memory sequencing reads of deoxyribonucleic acid (DNA) molecules of a cell-free bodily fluid sample of a subject; executing code with a computer processor to perform the following steps: generating from the sequence reads a first data set comprising for each genetic locus in a plurality of genetic loci a quantitative measure related to sequencing read coverage ("read coverage"); correcting the first data set by performing saturation equilibrium correction and probe efficiency correction; determining a baseline read coverage for the first data set, wherein the baseline read coverage relates to saturation equilibrium and probe efficiency; and determining a copy number state for each genetic locus in the plurality of genetic loci relative to the baseline read coverage.

[0015] In another aspect, the present disclosure provides a system comprising: a network; a database comprising computer memory configured to store nucleic acid (e.g., DNA) sequence data which are connected to the network; a bioinformatics computer comprising a computer memory and one or more computer processors, which computer is connected to the network;

wherein the computer further comprises machine-executable code which, when executed by the one or more computer processors, copies nucleic acid (e.g., DNA) sequence data stored on the database, writes the copied data to memory in the bioinformatics computer and performs steps including: generating from the nucleic acid (e.g., DNA) sequence data a first data set comprising for each genetic locus in a plurality of genetic loci a quantitative measure related to sequencing read coverage ("read coverage"); correcting the first data set by performing saturation equilibrium correction and probe efficiency correction; determining a baseline read coverage for the first data set, wherein the baseline read coverage relates to saturation equilibrium and probe efficiency; and determining a copy number state for each genetic locus in the plurality of genetic loci relative to the baseline read coverage. In some embodiments, the database is connected to a DNA sequencer.

INCORPORATION BY REFERENCE

[0016] All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings of which:

[0018] FIG. 1 illustrates exemplary oncogenes and targets for sequence capture probes.

[0019] FIG. 2 illustrates gene-level signal versus theoretical copy number across three spike-in and probe-level signal variation across spike-in genes

[0020] FIG. 3 illustrates a bait optimization experiment relating bait amount with unique molecular counts.

[0021] FIG. 4A and **FIG. 4B** illustrate the nonlinear effects of p (**FIG. 4A**) and GC content (**FIG. 4B**) on unique molecular counts.

[0022] FIG. 5 illustrates unique molecular counts per probe without saturation or probe-efficiency correction being performed.

[0023] FIG. 6 illustrates post-saturation correction unique molecular counts per probe.

[0024] FIG. 7 illustrates post-saturation and post-probe-efficiency corrected unique molecular counts per probe.

[0025] **FIG. 8** illustrates a proposed Langmuir model of interactions between true copy number and unique molecular counts related to probe saturation and probe efficiency.

[0026] **FIG. 9** illustrates the probe signal-noise reduction for the baselining genetic loci after saturation correct, probe efficiency correction, and a second round of probe efficiency correction in a typical clinical sample.

[0027] **FIG. 10A** and **FIG. 10B** illustrate post-saturation corrected UMCs plotted against the probe efficiencies determined in the reference sample in order to perform probe-efficiency correction. **FIG. 10A** is from a subject without copy number alteration in tumor cells. **FIG. 10B** is from a subject with copy number alteration in tumor cells.

[0028] **FIG. 11** illustrates a final report of saturation and probe efficiency corrected copy number variation detection in a patient sample. Stars above a sample indicate gene amplification detected based on the corrected signal and minor-allele frequency corrected baseline optimization.

[0029] **FIG. 12** illustrates a computer system 1201 that is programmed or otherwise configured to implement methods of the present disclosure.

[0030] **FIG. 13** illustrates observed copy number (CN) vs. theoretical CN for the gene ERBB2 as measured using a method of the present disclosure. Solid dots represent an observed copy number of ~2 (a diploid sample), open dots represent detected amplification events and the thick horizontal dashed line marks the mean gene CN cutoff.

[0031] **FIG. 14** illustrates observed copy number (CN) vs. theoretical CN for the gene ERBB2 as measured using a method of the present disclosure (dots) as compared to a control method (squares). Solid dots represent an observed copy number of ~2 (a diploid sample), open dots represent detected amplification events and the thick horizontal dashed line marks the mean gene CN cutoff.

[0032] **FIG. 15** illustrates probe copy number as plotted against probes used in a validation study for a method of the present disclosure (triangles) vs. a control method (X's).

DETAILED DESCRIPTION

Definitions

[0033] The term "genetic variant," as used herein, generally refers to an alteration, variant or polymorphism in a nucleic acid sample or genome of a subject. Such alteration, variant or polymorphism can be with respect to a reference genome, which may be a reference genome of the subject or other individual. Single nucleotide polymorphisms (SNPs) are a form of polymorphisms. In some examples, one or more polymorphisms comprise one or more single nucleotide variations (SNVs), insertions, deletions, repeats, small insertions, small deletions,

small repeats, structural variant junctions, variable length tandem repeats, and/or flanking sequences,. Copy number variants (CNVs), transversions and other rearrangements are also forms of genetic variation. A genomic alternation may be a base change, insertion, deletion, repeat, copy number variation, or transversion.

[0034] The term "polynucleotide," as used herein, generally refers to a molecule comprising one or more nucleic acid subunits. A polynucleotide can include one or more subunits selected from adenosine (A), cytosine (C), guanine (G), thymine (T) and uracil (U), or variants thereof. A nucleotide can include A, C, G, T or U, or variants thereof. A nucleotide can include any subunit that can be incorporated into a growing nucleic acid strand. Such subunit can be an A, C, G, T, or U, or any other subunit that is specific to one or more complementary A, C, G, T or U, or complementary to a purine (i.e., A or G, or variant thereof) or a pyrimidine (i.e., C, T or U, or variant thereof). A subunit can enable individual nucleic acid bases or groups of bases (e.g., AA, TA, AT, GC, CG, CT, TC, GT, TG, AC, CA, or uracil-counterparts thereof) to be resolved. In some examples, a polynucleotide is deoxyribonucleic acid (DNA) or ribonucleic acid (RNA), or derivatives thereof. A polynucleotide can be single-stranded or double stranded.

[0035] The term "subject," as used herein, generally refers to an animal, such as a mammalian species (e.g., human) or avian (e.g., bird) species, or other organism, such as a plant. More specifically, the subject can be a vertebrate, a mammal, a mouse, a primate, a simian or a human. Animals include, but are not limited to, farm animals, sport animals, and pets. A subject can be a healthy individual, an individual that has or is suspected of having a disease or a pre-disposition to the disease, or an individual that is in need of therapy or suspected of needing therapy. A subject can be a patient.

[0036] The term "genome" generally refers to an entirety of an organism's hereditary information. A genome can be encoded either in DNA or in RNA. A genome can comprise coding regions that code for proteins as well as non-coding regions. A genome can include the sequence of all chromosomes together in an organism. For example, the human genome has a total of 46 chromosomes. The sequence of all of these together constitutes a human genome.

[0037] The terms "adaptor(s)", "adaptor(s)" and "tag(s)" are used synonymously throughout this specification. An adaptor or tag can be coupled to a polynucleotide sequence to be "tagged" by any approach including ligation, hybridization, or other approaches.

[0038] The term "library adaptor" or "library adaptor" as used herein, generally refers to a molecule (e.g., a polynucleotide) whose identity (e.g., sequence) can be used to differentiate polynucleotides in a biological sample (also "sample" herein).

[0039] The term "sequencing adaptor," as used herein, generally refers to a molecule (e.g., a polynucleotide) that is adapted to permit a sequencing instrument to sequence a target

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

[0040] As used herein the terms "at least", "at most" or "about", when preceding a series, refers to each member of the series, unless otherwise identified.

[0041] The term "about" and its grammatical equivalents in relation to a reference numerical value can include a range of values up to plus or minus 10% from that value. For example, the amount "about 10" can include amounts from 9 to 11. In other embodiments, the term "about" in relation to a reference numerical value can include a range of values plus or minus 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, or 1% from that value.

[0042] The term "at least" and its grammatical equivalents in relation to a reference numerical value can include the reference numerical value and greater than that value. For example, the amount "at least 10" can include the value 10 and any numerical value above 10, such as 11, 100, and 1,000.

[0043] The term "at most" and its grammatical equivalents in relation to a reference numerical value can include the reference numerical value and less than that value. For example, the amount "at most 10" can include the value 10 and any numerical value under 10, such as 9, 8, 5, 1, 0.5, and 0.1.

[0044] The term "quantitative measure" refers to any measure of quantity including absolute and relative measures. A quantitative measure can be, for example, a number (e.g., a count), a percentage, a degree or a threshold.

[0045] The term "read coverage" refers to coverage by raw sequence reads or by processed sequence reads, such as unique molecular counts inferred from raw sequence reads.

[0046] The term "baseline read coverage" refers to expected read coverage of a probe in a sample comprising a diploid genome environment based on given probe parameters, such as GC content, probe efficiency, ligation efficiency, or pull down efficiency.

[0047] "Probe", as used herein, refers to a polynucleotide comprising a functionality. The functionality can be a detectable label (fluorescent), a binding moiety (biotin), or a solid support (a magnetically attractable particle or a chip).

[0048] "Complementarity" refers to the ability of a nucleic acid to form hydrogen bond(s) with another nucleic acid sequence by either traditional Watson-Crick or other non-traditional types. A percent complementarity indicates the percentage of residues in a nucleic acid molecule which can form hydrogen bonds (Watson-Crick base pairing) with a second nucleic acid sequence (5, 6, 7, 8, 9, 10 out of 10 being 50%, 60%, 70%, 80%, 90%, and 100% complementary, respectively). "Perfectly complementary" means that all the contiguous residues of a nucleic acid sequence will hydrogen bond with the same number of contiguous residues in a second nucleic acid sequence.

[0049] "Substantially complementary" as used herein refers to a degree of complementarity that is at least 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 97%, 98%, 99%, or 100% over a region of 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, or more nucleotides, or refers to two nucleic acids that hybridize under stringent conditions. Sequence identity, such as for the purpose of assessing percent complementarity, may be measured by any suitable alignment algorithm, including but not limited to the Needleman-Wunsch algorithm (see e.g. the EMBOSS Needle aligner available at the world wide web site: ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.html, optionally with default settings), the BLAST algorithm (see e.g. the BLAST alignment tool available at blast.ncbi.nlm.nih.gov/Blast.cgi, optionally with default settings), or the Smith-Waterman algorithm (see e.g. the EMBOSS Water aligner available at the world wide web site: ebi.ac.uk/Tools/psa/emboss_water/nucleotide.html, optionally with default settings). Optimal alignment may be assessed using any suitable parameters of a chosen algorithm, including default parameters.

[0050] "Hybridization" refers to a reaction in which one or more polynucleotides react to form a complex that is stabilized via hydrogen bonding between the bases of the nucleotide residues. The hydrogen bonding may occur by Watson Crick base pairing, Hoogstein binding, or in any other sequence specific manner according to base complementarity. The complex may comprise two strands forming a duplex structure, three or more strands forming a multi stranded complex, a single self-hybridizing strand, or any combination of these. A hybridization reaction may constitute a step in a more extensive process, such as the initiation of PCR, or the enzymatic

cleavage of a polynucleotide by an endonuclease. A second sequence that is complementary to a first sequence is referred to as the "complement" of the first sequence. The term "hybridizable" as applied to a polynucleotide refers to the ability of the polynucleotide to form a complex that is stabilized via hydrogen bonding between the bases of the nucleotide residues in a hybridization reaction.

[0051] The term "stringent hybridization conditions" refers to conditions under which a polynucleotide will hybridize preferentially to its target subsequence, and to a lesser extent to, or not at all to, other sequences. "Stringent hybridization" in the context of nucleic acid hybridization experiments are sequence dependent, and are different under different environmental parameters. An extensive guide to the hybridization of nucleic acids is found in Tijssen (1993) Laboratory Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Acid Probes part I chapter 2 "Overview of principles of hybridization and the strategy of nucleic acid probe assays", Elsevier, New York.

[0052] Generally, highly stringent hybridization and wash conditions are selected to be about 5° C lower than the thermal melting point (Tm) for the specific sequence at a defined ionic strength and pH. The Tm is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Very stringent conditions are selected to be equal to the Tm for a particular probe.

[0053] Stringent hybridization conditions include a buffer comprising water, a buffer (a phosphate, tris, SSPE or SSC buffer at pH 6-9 or pH 7-8), a salt (sodium or potassium), and a denaturant (SDS, formamide or tween) and a temperature of 37° C -70° C, 60° C -65° C.

[0054] An example of stringent hybridization conditions for hybridization of complementary nucleic acids which have more than 100 complementary residues on a filter in a Southern or northern blot is 50% formalin with 1 mg of heparin at 42° C, with the hybridization being carried out overnight. An example of highly stringent wash conditions is 0.15 M NaCl at 72° C for about 15 minutes. An example of stringent wash conditions is a 0.2X SSC wash at 65° C for 15 minutes (see, Sambrook et al. for a description of SSC buffer). Often, a high stringency wash is preceded by a low stringency wash to remove background probe signal. An example medium stringency wash for a duplex of, more than 100 nucleotides, is 1x SSC at 45° C for 15 minutes. An example low stringency wash for a duplex of, e. g., more than 100 nucleotides, is 4-6x SSC at 40° C for 15 minutes. In general, a signal to noise ratio of 2x (or higher) than that observed for an unrelated probe in the particular hybridization assay indicates detection of a specific hybridization.

[0055] In one aspect, the present disclosure provides a method comprising: (a) obtaining sequencing reads derived from deoxyribonucleic acid (DNA) molecules of a cell-free bodily

fluid sample of a subject; (b) generating a first data set, the first data set comprising, for each genetic locus in a plurality of genetic loci, a quantitative measure related to (i) guanine-cytosine content of the genetic locus and (ii) a quantitative measure related to sequencing read coverage of the genetic locus from the sequencing reads; (c) transforming the first data set into a second data set by: (i) removing from the first data set genetic loci that are high-variance genetic loci with respect to the quantitative measure related to sequencing read coverage, thereby providing a first set of remaining genetic loci; (ii) determining for each genetic locus from the first set of remaining genetic loci a quantitative measure related to the probability that a strand of DNA from the sample derived from the genetic locus is represented within the sequencing reads; (iii) determining a first transformation for the quantitative measure related to sequencing read coverage by relating the quantitative measure related to sequencing read coverage of the first set of remaining genetic loci to both the quantitative measure related to the GC content of the first set of remaining genetic loci and the quantitative measure related to the probability that a strand of DNA derived from the each locus in the first set of remaining genetic loci is represented within the sequencing reads; and (iv) applying the first transformation to the sequence read coverage of each genetic locus from the first set of remaining genetic loci to provide the second data set, wherein the second data set comprises a first set of transformed quantitative measures of sequencing read coverage of the first set of remaining genetic loci.

[0056] In some embodiments, the method further comprises transforming the second data set into a third data set by: (d) removing from the second data set genetic loci that are high-variance genetic loci with respect to the first set of transformed quantitative measures of sequencing read coverage, thereby providing a second set of remaining genetic loci; (e) determining a second transformation for the first set of transformed quantitative measures of sequencing read coverage related to the efficiency of the second set of remaining genetic loci; and (f) transforming the first set of transformed quantitative measures of sequencing read coverage of the second set of remaining genetic loci with the second transformation, thereby providing the third data set, wherein the third data set comprises a second set of transformed quantitative measures related to sequencing read coverage of the second set of remaining genetic loci of (d, i);

Obtaining sequencing reads from DNA molecules of a cell-free bodily fluid from a subject

[0057] Obtaining sequencing reads from DNA molecules of a cell-free bodily fluid of a subject can comprise obtaining a cell-free bodily fluid. Exemplary cell-free bodily fluids are or can be derived from serum, plasma, blood, saliva, urine, synovial fluid, whole blood, lymphatic fluid, ascites fluid, interstitial or extracellular fluid, the fluid in spaces between cells, including gingival crevicular fluid, bone marrow, cerebrospinal fluid, saliva, mucous, sputum, semen,

sweat, urine, or any other bodily fluids. A cell-free bodily fluid can be selected from the group consisting of plasma, urine, or cerebrospinal fluid. A cell-free bodily fluid can be plasma. A cell-free bodily fluid can be urine. A cell-free bodily fluid can be cerebrospinal fluid.

[0058] Nucleic acid molecules, including DNA molecules, can be extracted from cell-free bodily fluids. DNA molecules can be genomic DNA. DNA molecules can be from cells of healthy tissue of the subject. DNA molecules can be from noncancerous cells that have undergone somatic mutation. DNA molecules can be from a fetus in a maternal sample. The skilled worker will understand that, in embodiments wherein the DNA molecules are from a fetus in a maternal sample, a subject may refer to the fetus even though the sample is maternal. DNA molecules can be from precancerous cells of the subject. DNA molecules can be from cancerous cells of the subject. DNA molecules can be from cells within primary tumors of the subject. DNA molecules can be from secondary tumors of the subject. DNA molecules can be circulating DNA. The circulating DNA can comprise circulating tumor DNA (ctDNA). DNA molecules can be double-stranded or single-stranded. Alternatively, DNA molecule can comprise a combination of a double-stranded portion and a single-stranded portion. DNA molecules do not have to be cell-free. In some cases, the DNA molecules can be isolated from a sample. For example, DNA molecules can be cell-free DNA isolated from a bodily fluid, e.g., serum or plasma.

[0059] A sample can comprise various amounts of genome equivalents of nucleic acid molecules. For example, a sample of about 30 ng DNA can contain about 10,000 haploid human genome equivalents and, in the case of cfDNA, about 200 billion individual polynucleotide molecules. Similarly, a sample of about 100 ng of DNA can contain about 30,000 haploid human genome equivalents and, in the case of cfDNA, about 600 billion individual molecules.

[0060] Cell-free DNA molecules may be isolated and extracted from bodily fluids using a variety of techniques known in the art. In some cases, cell-free nucleic acids may be isolated, extracted and prepared using commercially available kits such as the Qiagen Qiamp® Circulating Nucleic Acid Kit protocol. In other examples, Qiagen Qubit™ dsDNA HS Assay kit protocol, Agilent™ DNA 1000 kit, or TruSeq™ Sequencing Library Preparation; Low-Throughput (LT) protocol may be used to quantify nucleic acids. Cell-free nucleic acids may be fetal in origin (via fluid taken from a pregnant subject), or may be derived from tissue of the subject itself. Cell-free nucleic acids can be derived from a neoplasm (e.g. a tumor or an adenoma).

[0061] Generally, cell-free nucleic acids are extracted and isolated from bodily fluids through a partitioning step in which cell-free nucleic acids, as found in solution, are separated from cells

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

[0062] Cell-free DNA molecules can be at most 500 nucleotides in length, at most 400 nucleotides in length, at most 300 nucleotides in length, at most 250 nucleotides in length, at most 225 nucleotides in length, at most 200 nucleotides in length, at most 190 nucleotides in length, at most 180 nucleotides in length, at most 170 nucleotides in length, at most 160 nucleotides in length, at most 150 nucleotides in length, at most 140 nucleotides in length, at most 130 nucleotides in length, at most 120 nucleotides in length, at most 110 nucleotides in length, or at most 100 nucleotides in length.

[0063] Cell-free DNA molecules can be at least 500 nucleotides in length, at least 400 nucleotides in length, at least 300 nucleotides in length, at least 250 nucleotides in length, at least 225 nucleotides in length, at least 200 nucleotides in length, at least 190 nucleotides in length, at least 180 nucleotides in length, at least 170 nucleotides in length, at least 160 nucleotides in length, at least 150 nucleotides in length, at least 140 nucleotides in length, at least 130 nucleotides in length, at least 120 nucleotides in length, at least 110 nucleotides in length, or at least 100 nucleotides in length. In particular, cell-free nucleic acids can be between 140 and 180 nucleotides in length.

[0064] Cell-free DNA can comprise DNA molecules from healthy tissue and tumors in various amounts. Tumor-derived cell-free DNA can be at least 0.1% of the total amount of cell-free DNA in the sample, at least 0.2% of the total amount of cell-free DNA in the sample, at least 0.5% of the total amount of cell-free DNA in the sample, at least 0.7% of the total amount of cell-free DNA in the sample, at least 1% of the total amount of cell-free DNA in the sample, at least 2% of the total amount of cell-free DNA in the sample, at least 3% of the total amount of cell-free DNA in the sample, at least 4% of the total amount of cell-free DNA in the sample, at least 5% of the total amount of cell-free DNA in the sample, at least 10% of the total amount of cell-free DNA in the sample, at least 15% of the total amount of cell-free DNA in the sample, at least 20% of the total amount of cell-free DNA in the sample, at least 25% of the total amount of

cell-free DNA in the sample, or at least 30% of the total amount of cell-free DNA in the sample, or more.

[0065] In some cases, DNA molecules can be sheared during the extraction process and comprise fragments between 100 and 400 nucleotides in length. In some cases, nucleic acids can be sheared after extraction can comprise nucleotides between 100 and 400 nucleotides in length. In some cases, DNA molecules are already between 100 and 400 nucleotides in length and additional shearing is not purposefully implemented.

[0066] A subject can be an animal. A subject can be a mammal, such as a dog, horse, cat, mouse, rat, or human. A subject can be a human. A subject can be suspected of having cancer. A subject can have previously received a cancer diagnosis. The cancer status of a subject may be unknown. A subject can be male or female. A subject can be at least 20 years old, at least 30 years old, at least 40 years old, at least 50 years old, at least 60 years old, or at least 70 years old.

[0067] Sequencing may be by any method known in the art. For example, sequencing techniques include classic techniques (e.g., dideoxy sequencing reactions (Sanger method) using labeled terminators or primers and gel separation in slab or capillary) and next generation techniques. Exemplary techniques include sequencing by synthesis using reversibly terminated labeled nucleotides, pyrosequencing, 454 sequencing, Illumina/Solexa sequencing, allele specific hybridization to a library of labeled oligonucleotide probes, sequencing by synthesis using allele specific hybridization to a library of labeled clones that is followed by ligation, real time monitoring of the incorporation of labeled nucleotides during a polymerization step, polony sequencing, SOLiD sequencing targeted sequencing, single molecule real-time sequencing, exon sequencing, electron microscopy-based sequencing, panel sequencing, transistor-mediated sequencing, direct sequencing, random shotgun sequencing, whole-genome sequencing, sequencing by hybridization, , capillary electrophoresis, gel electrophoresis, duplex sequencing, cycle sequencing, single-base extension sequencing, solid-phase sequencing, high-throughput sequencing, massively parallel signature sequencing, emulsion PCR, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, sequencing by reversible dye terminator, paired-end sequencing, near-term sequencing, exonuclease sequencing, sequencing by ligation, short-read sequencing, single-molecule sequencing, real-time sequencing, reverse-terminator sequencing, nanopore sequencing, MS-PET sequencing, and a combination thereof. In some embodiments, the sequencing method is massively parallel sequencing, that is, simultaneously (or in rapid succession) sequencing any of at least 100, 1000, 10,000, 100,000, 1 million, 10 million, 100 million, or 1 billion polynucleotide molecules. In some embodiments, sequencing can be performed by a gene analyzer such as, for example, gene analyzers commercially available from Illumina or Applied Biosystems. Sequencing of separated

molecules has more recently been demonstrated by sequential or single extension reactions using polymerases or ligases as well as by single or sequential differential hybridizations with libraries of probes. Sequencing may be performed by a DNA sequencer (e.g., a machine designed to perform sequencing reactions). In some embodiments, a DNA sequencer can comprise or be connected to a database, for example, that contains DNA sequence data.

[0068] A sequencing technique that can be used includes, for example, use of sequencing-by-synthesis systems. In the first step, DNA is sheared into fragments of approximately 300-800 base pairs, and the fragments are blunt ended. Oligonucleotide adaptors are then ligated to the ends of the fragments. The adaptors serve as primers for amplification and sequencing of the fragments. The fragments can be attached to DNA capture beads, e.g., streptavidin-coated beads using, e.g., Adaptor B, which contains 5'-biotin tag. The fragments attached to the beads are PCR amplified within droplets of an oil-water emulsion. The result is multiple copies of clonally amplified DNA fragments on each bead. In the second step, the beads are captured in wells (pico-liter sized). Pyrosequencing is performed on each DNA fragment in parallel. Addition of one or more nucleotides generates a light signal that is recorded by a CCD camera in a sequencing instrument. The signal strength is proportional to the number of nucleotides incorporated. Pyrosequencing makes use of pyrophosphate (PPi) which is released upon nucleotide addition. PPi is converted to ATP by ATP sulfurylase in the presence of adenosine 5' phosphosulfate. Luciferase uses ATP to convert luciferin to oxyluciferin, and this reaction generates light that is detected and analyzed.

[0069] Another example of a DNA sequencing technique that can be used is SOLiD technology by Applied Biosystems from Life Technologies Corporation (Carlsbad, Calif.). In SOLiD sequencing, genomic DNA is sheared into fragments, and adaptors are attached to the 5' and 3' ends of the fragments to generate a fragment library. Alternatively, internal adaptors can be introduced by ligating adaptors to the 5' and 3' ends of the fragments, circularizing the fragments, digesting the circularized fragment to generate an internal adaptor, and attaching adaptors to the 5' and 3' ends of the resulting fragments to generate a mate-paired library. Next, clonal bead populations are prepared in microreactors containing beads, primers, template, and PCR components. Following PCR, the templates are denatured and beads are enriched to separate the beads with extended templates. Templates on the selected beads are subjected to a 3' modification that permits bonding to a glass slide. The sequence can be determined by sequential hybridization and ligation of partially random oligonucleotides with a central determined base (or pair of bases) that is identified by a specific fluorophore. After a color is recorded, the ligated oligonucleotide is removed and the process is then repeated.

[0070] Another example of a DNA sequencing technique that can be used is ion semiconductor sequencing using, for example, a system sold under the trademark ION TORRENT by Ion Torrent by Life Technologies (South San Francisco, Calif.). Ion semiconductor sequencing is described, for example, in Rothberg, et al., An integrated semiconductor device enabling non-optical genome sequencing, Nature 475:348-352 (2011); U.S. Pub. 2010/0304982; U.S. Pub. 2010/0301398; U.S. Pub. 2010/0300895; U.S. Pub. 2010/0300559; and U.S. Pub. 2009/0026082, the contents of each of which are incorporated by reference in their entirety.

[0071] Another example of a sequencing technology that can be used is Illumina sequencing. Illumina sequencing is based on the amplification of DNA on a solid surface using fold-back PCR and anchored primers. Genomic DNA is fragmented, and adapters are added to the 5' and 3' ends of the fragments. DNA fragments that are attached to the surface of flow cell channels are extended and bridge amplified. The fragments become double stranded, and the double stranded molecules are denatured. Multiple cycles of the solid-phase amplification followed by denaturation can create several million clusters of approximately 1,000 copies of single-stranded DNA molecules of the same template in each channel of the flow cell. Primers, DNA polymerase and four fluorophore-labeled, reversibly terminating nucleotides are used to perform sequential sequencing. After nucleotide incorporation, a laser is used to excite the fluorophores, and an image is captured and the identity of the first base is recorded. The 3' terminators and fluorophores from each incorporated base are removed and the incorporation, detection and identification steps are repeated. Sequencing according to this technology is described in U.S. Pat. No. 7,960,120; U.S. Pat. No. 7,835,871; U.S. Pat. No. 7,232,656; U.S. Pat. No. 7,598,035; U.S. Pat. No. 6,911,345; U.S. Pat. No. 6,833,246; U.S. Pat. No. 6,828,100; U.S. Pat. No. 6,306,597; U.S. Pat. No. 6,210,891; U.S. Pub. 2011/0009278; U.S. Pub. 2007/0114362; U.S. Pub. 2006/0292611; and U.S. Pub. 2006/0024681, each of which are incorporated by reference in their entirety.

[0072] Another example of a sequencing technology that can be used includes the single molecule, real-time (SMRT) technology of Pacific Biosciences (Menlo Park, Calif.). In SMRT, each of the four DNA bases is attached to one of four different fluorescent dyes. These dyes are phospholinked. A single DNA polymerase is immobilized with a single molecule of template single stranded DNA at the bottom of a zero-mode waveguide (ZMW). It takes several milliseconds to incorporate a nucleotide into a growing strand. During this time, the fluorescent label is excited and produces a fluorescent signal, and the fluorescent tag is cleaved off. Detection of the corresponding fluorescence of the dye indicates which base was incorporated. The process is repeated.

[0073] Another example of a sequencing technique that can be used is nanopore sequencing (Soni & Meller, 2007, Progress toward ultrafast DNA sequence using solid-state nanopores, Clin Chem 53(11):1996-2001). A nanopore is a small hole, of the order of 1 nanometer in diameter. Immersion of a nanopore in a conducting fluid and application of a potential across it results in a slight electrical current due to conduction of ions through the nanopore. The amount of current which flows is sensitive to the size of the nanopore. As a DNA molecule passes through a nanopore, each nucleotide on the DNA molecule obstructs the nanopore to a different degree. Thus, the change in the current passing through the nanopore as the DNA molecule passes through the nanopore represents a reading of the DNA sequence.

[0074] Another example of a sequencing technique that can be used involves using a chemical-sensitive field effect transistor (chemFET) array to sequence DNA (for example, as described in U.S. Pub. 2009/0026082). In one example of the technique, DNA molecules can be placed into reaction chambers, and the template molecules can be hybridized to a sequencing primer bound to a polymerase. Incorporation of one or more triphosphates into a new nucleic acid strand at the 3' end of the sequencing primer can be detected by a change in current by a chemFET. An array can have multiple chemFET sensors. In another example, single nucleic acids can be attached to beads, and the nucleic acids can be amplified on the bead, and the individual beads can be transferred to individual reaction chambers on a chemFET array, with each chamber having a chemFET sensor, and the nucleic acids can be sequenced.

[0075] Another example of a sequencing technique that can be used involves using an electron microscope as described, for example, by Moudrianakis, E. N. and Beer M., in Base sequence determination in nucleic acids with the electron microscope, III. Chemistry and microscopy of guanine-labeled DNA, PNAS 53:564-71 (1965). In one example of the technique, individual DNA molecules are labeled using metallic labels that are distinguishable using an electron microscope. These molecules are then stretched on a flat surface and imaged using an electron microscope to measure sequences.

[0076] Prior to sequencing, adaptor sequences can be attached to the nucleic acid molecules and the nucleic acids can be enriched for particular sequences of interest. Sequence enrichment can occur before or after the attachment of adaptor sequence.

[0077] The nucleic acid molecules or enriched nucleic acid molecules can be attached to any sequencing adaptor suitable for use on any sequencing platform disclosed herein. For example, a sequence adaptor can comprise a flow cell sequence, a sample barcode, or both. In another example, a sequence adaptor can be a hairpin shaped adaptor, a Y-shaped adaptor, a forked adaptor, and/or comprise a sample barcode. In some cases, the adaptor does not comprise a sequencing primer region. In some cases the adaptor-attached DNA molecules are amplified,

and the amplification products are enriched for specific sequences as described herein. In some cases, the DNA molecules are enriched for specific sequences after preparing a sequencing library. Adaptors can comprise barcode sequence. The different barcode can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or more (or any length as described throughout) nucleic acid bases, e.g., 7 bases. The barcodes can be random sequences, degenerate sequences, semi-degenerate sequences, or defined sequences. In some cases, there is a sufficient diversity of barcodes that substantively (e.g., at least 70%, at least 80%, at least 90%, or at least 99% of) each nucleic acid molecule is tagged with a different barcode sequence. In some cases, there is a sufficient diversity of barcodes that substantively (e.g., at least 70%, at least 80%, at least 90%, or at least 99% of) each nucleic acid molecule from a particular genetic locus is tagged with a different barcode sequence.

[0078] A sequencing adaptor can comprise a sequence capable of hybridizing to one or more sequencing primers. A sequencing adaptor can further comprise a sequence hybridizing to a solid support, e.g., a flow cell sequence. For example, a sequencing adaptor can be a flow cell adaptor. The sequencing adaptors can be attached to one or both ends of a polynucleotide fragment. In another example, a sequencing adaptor can be hairpin shaped. For example, the hairpin shaped adaptor can comprise a complementary double-stranded portion and a loop portion, where the double-stranded portion can be attached (e.g., ligated) to a double-stranded polynucleotide. Hairpin shaped sequencing adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times.

[0079] In some cases, none of the library adaptors contains a sample identification motif (or sample molecular barcode). Such sample identification motif can be provided via sequencing adaptors. A sample identification motif can include a sequencer of at least 4, 5, 6, 7, 8, 9, 10, 20, 30, or 40 nucleotide bases that permits the identification of polynucleotide molecules from a given sample from polynucleotide molecules from other samples. For example, this can permit polynucleotide molecules from two subjects to be sequenced in the same pool and sequence reads for the subjects subsequently identified.

[0080] A sequencer motif includes nucleotide sequence(s) needed to couple a library adaptor to a sequencing system and sequence a target polynucleotide coupled to the library adaptor. The sequencer motif can include a sequence that is complementary to a flow cell sequence and a sequence (sequencing initiation sequence) that can be selectively hybridized to a primer (or priming sequence) for use in sequencing. For example, such sequencing initiation sequence can be complementary to a primer that is employed for use in sequence by synthesis (e.g., Illumina).

Such primer can be included in a sequencing adaptor. A sequencing initiation sequence can be a primer hybridization site.

[0081] In some cases, none of the library adaptors contains a complete sequencer motif. The library adaptors can contain partial or no sequencer motifs. In some cases, the library adaptors include a sequencing initiation sequence. The library adaptors can include a sequencing initiation sequence but no flow cell sequence. The sequence initiation sequence can be complementary to a primer for sequencing. The primer can be a sequence specific primer or a universal primer. Such sequencing initiation sequences may be situated on single-stranded portions of the library adaptors. As an alternative, such sequencing initiation sequences may be priming sites (e.g., kinks or nicks) to permit a polymerase to couple to the library adaptors during sequencing.

[0082] Adaptors can be attached to DNA molecules by ligation. In some cases, the adaptors are ligated to duplex DNA molecules such that each adaptor differently tags complementary strands of the DNA molecule. In some cases, adaptor sequences can be attached by PCR, wherein a first portion of a single-stranded DNA is complementary to a target sequence and a second portion comprises the adaptor sequence.

[0083] Enrichment for particular sequences of interest can be performed by sequence capture methods. Sequence capture can be performed using immobilized probes that hybridize to the targets of interest. Sequence capture can be performed using probes attached to functional groups, e.g., biotin, that allow probes hybridized to specific sequences to be enriched for from a sample by pulldown. In some cases, prior to hybridization to functionalized probes, specific sequences such as adaptor sequences from library fragments can be masked by annealing complementary, non-functionalized polynucleotide sequences to the fragments in order to reduce non-specific or off-target binding. Sequence probes can target specific genes. Sequence capture probes can target specific genetic loci or genes. Such genes can be oncogenes. Exemplary genes targeted by capture probes include those shown in **FIG. 1**. Exemplary genes with point mutations (SNVs) include, but are not limited to, AKT1, ATM, CCNE1, CTNNB1, FGFR1, GNAS, JAK3, MLH1, NPM1, PTPN11, RIT1, TERT, ALK, BRAF, CDH1, EGFR, FGFR2, HNF1A, KIT MPL, NRAS, RAF1, ROS1, TP53, APC, BRCA1, CDK4, ERBB2, FGFR3, HRAS, KRAS, MYC, NTRK1, RB1, SMAD4, TSC1, AR, BRCA2, CDK6, ESR1, GATA3, IDH2, MAP2K2, NFE2L2, PIK3CA, RHEB, SRC, ARID1A, CCND2, CDKN2B, FBXW7, GNAQ, JAK2, MET, NOTCH1, PTEN, RHOA, and STK11. Exemplary genes with copy number variations include, but are not limited to, AR, CCNE1, CDK6, ERBB2, FGFR2, KRAS, MYC, PIK3CA, BRAF, CDK4, EGFR, FGFR1, KIT, MET, PDGFRA, and RAF1. Exemplary genes with gene fusions include, but are not limited to: ALK, FGFR2, FGFR3,

NTRK1, RET, and ROS1. Exemplary genes with indels include, but are not limited to: EGFR (for example, at exons 19 and 20), ERBB2 (for example, at exons 19 and 20), and MET (for example, skipping exon 14). Exemplary targets can include CCND1 and CCND2. Sequence capture probes can tile across a gene (e.g., probes can target overlapping regions). Sequence probes can target non-overlapping regions. Sequence probes can be optimized for length, melting temperature, and secondary structure.

Quantitative Measures of Guanine-Cytosine (GC) Content

[0084] Guanine-cytosine content is the percentage of nitrogenous bases of a DNA molecule that are either guanine or cytosine. A quantitative measure related to GC content for a genetic locus can be the GC content of the entire genetic locus. A quantitative measure related to GC content for a genetic locus can be the GC content of the exonic regions of the gene. A quantitative measure related to GC content for a genetic locus can be the GC content of the regions covered by reads mapping to the genetic locus. A quantitative measure related to GC content can be the GC content of the sequence capture probes corresponding to the genetic locus. A quantitative measure related to GC content for a genetic locus can be a measure related to central tendency of the GC content of the sequence capture probes corresponding to the genetic locus. The measure related to central tendency can be any measure of central tendency such as mean, median, or mode. The measure related to central tendency can be the median. GC content of a given region can be measured by dividing the number of guanosine and cytosine bases by the total number of bases over that region.

Quantitative Measures of Sequencing Read Coverage

[0085] A quantitative measure related to sequencing read coverage is a measure indicative of the number of reads derived from a DNA molecule corresponding to a genetic locus (e.g., a particular position, base, region, gene or chromosome from a reference genome). In order to associate reads to a genetic locus, the reads can be mapped or aligned to the reference. Software to perform mapping or aligning (e.g., Bowtie, BWA, mrsFAST, BLAST, BLAT) can associate a sequencing read with a genetic locus. During the mapping process, particular parameters can be optimized. Non-limiting examples of optimization of the mapping processing can include masking repetitive regions; employing mapping quality (e.g., MAPQ) score cut-offs; using different seed lengths to generate alignments; and limiting the edit distance between positions of the genome.

[0086] Quantitative measures associated with sequencing read coverage can include counts of reads associated with a genetic locus. In some cases, the counts are transformed into new metrics

to mitigate the effects of differing sequencing depth, library complexity, or size of the genetic locus. Exemplary metrics are Read Per Kilobase per Million (RPKM), Fragments Per Kilobase per Million (FPKM), Trimmed Mean of M values (TMM), variance stabilized raw counts, and log transformed raw counts. Other transformations are also known to those of skill in the art that may be used for particular applications.

[0087] Quantitative measures can be determined using collapsed reads, wherein each collapsed read corresponds to an initial template DNA molecule. Methods to collapse and quantify read families are found in PCT/US2013/058061 and PCT/US2014/000048, each of which is herein incorporated by reference in its entirety. In particular, collapsing methods can be employed that use barcodes and sequence information from the sequencing read to collapse reads into families, such that each family shares barcode sequences and at least a portion of the sequencing read sequence. Each family is then, for the majority of the families, derived from a single initial template DNA molecule. Counts derived from mapping sequences from families can be referred to as "unique molecular counts" (UMCs). In some cases, determining a quantitative measure related to sequencing read coverage comprises normalizing UMCs by a metric related to library size to provide normalized UMCs ("normalized UMCs"). Exemplary methods are dividing the UMC of a genetic locus by the sum of all UMCs; dividing the UMC of a genetic locus by the sum of all autosomal UMCs. When comparing multiple sequencing read data sets, UMCs can, for example, be normalized by the median UMCs of the genetic loci of the two sequencing read data sets. In some cases, the quantitative measure related to sequencing read coverage can be normalized UMCs that are further normalized as follows: (i) normalized UMCs are determined for corresponding genetic loci from sequencing reads derived from training samples; (ii) for each genetic locus, normalized UMCs of the sample are normalized by the median of the normalized UMCs of the training samples at the corresponding loci, thereby providing Relative Abundances (RAs) of genetic loci.

[0088] Consensus sequences can identified based on their sequences, for example by collapsing sequencing reads based on identical sequences within the first 5, 10, 15, 20, or 25 bases. In some cases, collapsing allows for 1 difference, 2 differences, 3 differences, 4 differences, or 5 differences in the reads that are otherwise identical. In some cases, collapsing uses the mapping position of the read, for example the mapping position of the initial base of the sequencing read. In some cases, collapsing uses barcodes, and sequencing reads that share barcode sequences are collapsed into a consensus sequence. In some cases, collapsing uses both barcodes and the sequence of the initial template molecules. For example, all reads that share a barcode and map to the same position in the reference genome can be collapsed. In another example, all reads that

share a barcode and a sequence of the initial template molecule (or a percentage identity to a sequence of the initial template molecule) can be collapsed.

[0089] In some cases, quantitative measures of sequencing read coverage are determined for specific sub-regions of a genome. Regions can be bins, genes of interest, exons, regions corresponding to sequence probes, regions corresponding to primer amplification products, or regions corresponding to primer binding sites. In some cases, sub-regions of the genome are regions corresponding to sequence capture probes. A read can map to a region corresponding to the sequence capture probe if at least a portion of the read maps at least a portion of the region corresponding to the sequence capture probe. A read can map to a region corresponding to the sequence capture probe if at least a portion of the read maps to the majority of the region corresponding to the sequence capture probe. A read can map to a region corresponding to the sequence capture probe if at least a portion of the read maps across the center point of the region corresponding to the sequence capture probe. In some cases, a quantitative measure related to sequencing read coverage of a genetic locus is the median of the RAs of the probes corresponding to genomic locations within the genetic locus. For example, if KRAS is covered by three probes, which have RAs of 2, 3, and 5, the RA of the genetic locus would be 3.

"Saturation equilibrium" correction

[0090] In general, the methods described herein can be used to increase the specificity and sensitivity of variant calling (e.g., detecting copy number variants) in a nucleic acid sample. For example, the methods can decrease the amount of noise or distortion in a data sample, reducing the number of false positive variants detected. As noise and/or distortion decrease, specificity and sensitivity increase. Noise can be thought of as an unwanted random addition to a signal. Distortion can be thought of as an alteration in the amplitude of a signal or portion of a signal.

[0091] Noise can be introduced through errors in copying and/or reading a polynucleotide. For example, in a sequencing process, a single polynucleotide can first be subject to amplification. Amplification can introduce errors, so that a subset of the amplified polynucleotides may contain, at a particular locus, a base that is not the same as the original base at that locus. Furthermore, in the reading process a base at any particular locus may be read incorrectly. As a consequence, the collection of sequence reads can include a certain percentage of base calls at a locus that are not the same as the original base. In typical sequencing technologies this error rate can be in the single digits, e.g., 2%-3%. When a collection of molecules that are all presumed to have the same sequence are sequenced, this noise is sufficiently small that one can identify the original base with high reliability.

[0092] However, if a collection of parent polynucleotides includes a subset of polynucleotides having sequence variants at a particular locus, noise can be a significant problem. This can be the case, for example, when cell free DNA includes not only germline DNA, but DNA from another source, such as fetal DNA or DNA from a cancer cell. In this case, if the frequency of molecules with sequence variants is in the same range as the frequency of errors introduced by the sequencing process, then true sequence variants may not be distinguishable from noise. This could interfere, for example, with detecting sequence variants in a sample.

[0093] Distortion can be manifested in the sequencing process as a difference in signal strength, e.g., total number of sequence reads, produced by molecules in a parent population at the same frequency. Distortion can be introduced, for example, through amplification bias, GC bias, or sequencing bias. This could interfere with detecting copy number variation in a sample. GC bias results in the uneven representation of areas rich or poor in GC content in the sequence reading.

[0094] Methods disclosed herein comprise determining an initial set of genetic loci for use in determining a baseline by removing from a data set those genetic loci for which the quantitative measure related to sequencing read coverage or the transformed quantitative measure related to sequencing read coverage differs most from a predictive model (which can be referred to herein as removing high-variance genetic loci), thereby providing a first set of remaining genetic loci. In some instances, removing these genetic loci comprises fitting a model that relates the quantitative measures related to sequencing read coverage to the quantitative measures related to GC content of the genetic loci. For example, the predictive model can relate the RAs of the genetic loci to the GC content of the loci. In some cases, the predictive model is a regression model, including non-parametric regression models such as LOESS and LOWESS regression models. In some cases, baselining is performed by removing 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, or 70% of the genetic loci that deviate the most from the predictive model. In some cases, baselining is performed by removing at least 5%, at least 10%, at least 15%, at least 20%, at least 25%, at least 30%, at least 35%, at least 40%, at least 45%, at least 50%, at least 55%, at least 60%, at least 65%, or at least 70% of the genetic loci that deviate the most from the predictive model. In some cases, deviation is determined by measuring the residuals of the genetic loci relative to the model. The exact cut-off can be chosen to provide exclude a specific amount of variance from the remaining genetic loci.

[0095] Methods to determine a quantitative measure related to the probability that a strand of DNA from the sample derived from the genetic locus is represented within the sequencing reads are disclosed in PCT/US2014/072383, which is hereby incorporated by reference in its entirety. Determining the quantitative measure can comprise estimating number of initial template DNA molecules derived from a locus that were present in the sample. The probability that a double

strand polynucleotide generates no sequence reads can be determined based on the relative number of reads representing both strands of an initial template DNA molecule and reads representing only a single strand of an initial template DNA molecule.

[0096] The number of undetected initial template DNA molecules in a sample can be estimated based on the relative number of reads representing both strands of an initial template DNA molecule and reads representing only a single strand of an initial template DNA molecule. As an example, counts for a particular genetic locus, Locus A, are recorded, where 1000 molecules are paired (e.g., both strands are detected) and 1000 molecules are unpaired (e.g., only a single strand is detected). It should be noted that the terms "paired" and "unpaired" as used herein are distinct from these terms as sometimes applied to sequencing reads to indicated whether both ends or a single-end of a molecule are sequenced. Assuming a uniform probability, p, for an individual Watson or Crick strand to make it through the process subsequent to conversion, one can calculate the proportion of molecules that fail to make it through the process (Unseen) as follows: R, the ratio of paired to unpaired molecules = 1000/1000 = 1, therefore $R=1=p^2/(2p(1-p))$. This implies that p =2/3 and that the quantity of lost molecules is equal to $(1-p)^2 = 1/9$. Thus in this example, approximately 11% of converted molecules are lost and never detected. In addition to using binomial distribution, other methods of estimating numbers of unseen molecules include exponential, beta, gamma or empirical distributions based on the redundancy of sequence reads observed. In the latter case, the distribution of read counts for paired and unpaired molecules can be derived from such redundancy to infer the underlying distribution of original polynucleotide molecules at a particular locus. This can often lead to a better estimation of the number of unseen molecules. In some cases, p is the quantitative measure related to the probability that a strand of DNA from the sample derived from the genetic locus is represented in the sequencing reads. In some cases, p is similarly derived, but a different model of read distribution is used (e.g., binomial, poisson, beta, gamma, and negative binomial distribution).

[0097] A transformation for the quantitative measure related to sequencing read coverage can be determined by relating the quantitative measure or transformed sequencing read coverage from a set of genetic loci with high-variance genetic loci removed to the quantitative measure related to GC content and the quantitative measure related to the probability that a strand of DNA derived from the genetic locus is represented within the sequencing reads. In some cases, the remaining genetic loci are assumed to be diploid and/or to be present at the same copy number. In some instances, a transformation is determined by fitting a measure related to central tendency of the quantitative measures related to sequencing read coverage of the remaining genetic loci by the quantitative measure related to GC content and the quantitative measure related to the probability that a strand of DNA derived from the genetic locus is represented within the

sequencing reads. A transformation can, for example, (i) fit the central tendency of the quantitative measures sequencing read coverage of the remaining genetic loci after removal of high-variance genetic loci by both the quantitative measures related to GC content and the quantitative measures related to the probability that a strand of DNA derived from the genetic locus is represented within the sequencing reads. In some instances, the measure related to central tendency of the quantitative measures of sequencing read coverage of the remaining loci is the central tendency of the UMCs of the remaining genetic loci. In some instances, a surface approximation is used to fit a surface of UMCs of the remaining genetic loci or the central tendency of the UMCs of the remaining genetic loci by (i) the quantitative measures related to GC content and (ii) the quantitative measures related to the probability that a strand of DNA derived from the genetic locus is represented within the sequencing reads. For example, the surface approximation can be a two-dimensional second-degree polynomial surface fit of the measure related to initial template DNA molecules (e.g., UMCs) by the quantitative measures of GC content and p. In some cases, the transformed quantitative measure related to sequencing coverage is the value expected based on the transformation determined above calculated from (i) the quantitative measures related to GC content and (ii) the quantitative measures related to the probability that a strand of DNA derived from the genetic locus is represented within the sequencing reads. In some cases, the transformed quantitative measure related to sequencing coverage is the residual of each genetic locus (e.g., the difference or quotient of the expected quantitative measure related to sequencing read coverage of a locus based on the surface approximation and the observed quantitative measure related to sequencing read coverage of the genetic locus in the sample). Optionally, after the transformed quantitative measure related to sequencing coverage is determined, high-variance genetic loci can again be removed as described above based on the new transformed quantitative measures of sequencing read coverage.

"Probe efficiency" correction

[0098] Disclosed herein are methods to determine and remove biases for genetic loci using reference samples. In some cases, the reference samples are sequencing reads from cell-free DNA from subjects without cancer. In some cases, the reference samples are sequencing reads from cell-free DNA from subjects with cancer cells that substantially lack copy number variation in the genetic loci of interest. In some cases, the reference samples are sequencing reads from cell-free DNA from subjects with cancer, where regions suspected of have undergone copy number variation are excluded from analysis. In some cases, the reference

sample is a plasma sample from a subject without cancer. In some cases, the reference sample is a plasma sample from a subject with cancer.

[0099] Each of the genetic loci of the reference samples can be processed as described above in "saturation equilibrium correction" to provide transformed quantitative measures of sequencing read coverage. In some cases, the transformed quantitative measure related to sequencing coverage is the value expected based on the transformation determined above calculated from (i) the quantitative measures related to GC content and (ii) the quantitative measures related to the probability that a strand of DNA derived from the genetic locus from the reference genetic loci is represented within the sequencing reads. In some cases, the transformed quantitative measure related to sequencing coverage is the residual of each reference genetic locus (e.g., the difference or quotient of the expected quantitative measure related to sequencing read coverage of a locus based on the surface approximation and the observed quantitative measure related to sequencing read coverage of the genetic locus in the reference sample). The transformed quantitative measure related to sequencing read coverage of the genetic locus in the reference sample can be thought of as the "efficiency" of the genetic locus. For example, a genetic locus that is inefficiently amplified will have a lower UMC than a genetic locus (present at the same copy number in the sample) that is very efficiently amplified.

[00100] The transformed quantitative measure related to sequencing read coverage of the sample can be corrected based on the determined efficiency of the genetic loci from the reference sample(s). This correction can reduce variance introduced into the sample by the process of producing the sequencing reads from the sample, which can be related to ligation efficiency, pulldown efficiency, PCR efficiency, flow cell clustering loss, demultiplexing loss, collapsing loss, and alignment loss. In one embodiment, correction comprises dividing or subtracting the post-saturation transformed quantitative measures of sequencing coverage of the sample by the predicted post-saturation transformed quantitative measure related to sequencing coverage. In some instances, the predicted post-saturation transformed quantitative measure related to sequencing coverage of the genetic loci is determined by fitting a relationship between the post-saturation transformed quantitative measure related to sequencing coverage of the genetic loci from the sample and the post-saturation transformed quantitative measure related to sequencing read coverage of the references. In some cases, fitting comprises performing local regression (e.g., LOESS or LOWESS) or robust linear regression of the post-saturation transformed quantitative measure related to sequencing coverage of the genetic loci from the sample on the post-saturation transformed quantitative measure related to sequencing read coverage of the references. In some cases, the fitting can be linear regression, non-linear regression, or non-parametric regression.

[00101] Optionally, the transformed quantitative measure from the probe efficiency correction can be the input into the "saturation equilibrium correction" transformation to produce a third, further transformed quantitative measure related to sequencing read coverage with reduced variance. In general, transformed quantitative measures of sequencing coverage can be transformed using any of the methods disclosed herein additional times in order to further reduce the variance within the transformed quantitative measures of sequencing read coverage.

Gene level summaries

[00102] Gene level summaries of inferred copy number can be determined based on the transformed quantitative measures of sequencing read coverage determined as disclosed herein. Copy number can be inferred relative to the baseline selected in the above operations by discarding high variance genetic loci. For example, if the remaining genetic loci are inferred to be diploid in the sample, then genetic loci for which the transformed quantitative measure related to sequencing coverage differ from the baseline can be inferred to have undergone copy number alteration in the tumor cells. In some instances, gene-level z-scores are calculated using observed gene-level median of probe signal and estimated standard deviation calculated using observed probe-level standard deviation estimate in a gene and whole-genome normal diploid probe signal standard deviation.

Minor-allele frequency baseline optimization

[00103] Provided herein are methods to detect errors and correct errors in gene level summaries of copy number described herein using minor allele frequencies of variants in the sequencing reads. Sequence variants present in between 10% and 90%, between 20% and 80%, between 30% and 70%, between 40% and 60%, or approximately 50% of sequencing reads from nucleic acids from a cell-free bodily fluid can be heterozygous variants present in the germline sequence of the subject. In some instances, genetic loci have been determined to have undergone amplification as described above. The quantities of variants are compared to the inferred copy number to determine if variant frequency is inconsistent with the inferred copy number. In one example, heterozygous genetic loci can be examined in the genetic loci that were used to determine the baseline copy number (e.g., the genetic loci remaining after exclusion of the high-variance genetic loci). In some cases, numerous genetic loci in the sample have been amplified, and this baseline can be misidentified. In such cases, heterozygosity may deviate from a 1:1 ratio, and the inaccurate baselining is detected and corrected. In a second example, example, a genetic locus can be inferred to be present at a triploid copy number based on the transformed quantitative measure related to sequencing read coverage. If the germline genome of the subject

has one chromosome with a first allele of the genetic locus and a second chromosome had a second allele, then the first or second allele may have duplicated in the cancer cells.

Langmuir-like saturation model

[00104] Without being bound by theory, disclosed herein is a Langmuir-like saturation model assumed to be the governing mechanism of bait-cfDNA interactions based on exploration of historical clinical data as well as targeted experiments involving synthetic spike-in model systems. Hence, in the absence of interfering assay effects (e.g. ligation efficiencies, PCR amplification biases, sequencing artifacts, etc), bait pulldown process may be described as

$$\text{Unique molecule count} = \mathrm{I}_{\mathrm{sat}} \frac{K \bullet CopyNumber}{1 + K \bullet CopyNumber}$$

[00105] K in this description is bait efficiency, which is dependent on bait sequence characteristics and its interactions with DNA fragments in genomic vicinity of the targeted bait location. I_{sat} is a saturation parameter driven by the limited initial bait count in the pulldown reaction, which is a function of total bait pool concentration as well as replication count. Replication count as used herein refers to the relative or absolute amount of sequence capture probe present. For example, sequence capture arrays can provide for different molar quantities of probes on an array to account for differing probe efficiencies. **FIG. 8** illustrates the model relating true copy number and unique molecule count based on bait efficiency, K, and saturation parameter I_{sat}.

[00106] Bait efficiency K is largely driven by GC content, while I_{sat} is driven by more complex bait exhaustion mechanisms and RNA secondary structure interactions that can be crudely examined by studying unique molecule count vs. total read count interactions. Aside from non-linear pulldown reactions, probe signal can be further modeled by a multiplicative model involving the following assumption: under a naive model cfDNA fragments are uniformly distributed by genomic position with stochastic sampling process being the dominating factor contributing to coverage variation. Then, copy number signal (e.g., UMCs) can be modeled by relating the observed UMC to the true molecular count in the sample, taking into account the effects the underlying positional cfDNA profile, ligation efficiency, pulldown efficiency, PCR efficiency, flowcell clustering loss, demultiplexing loss, collapsing loss, and alignment loss.

[00107] Aside from non-linear pulldown reaction, probe signal can be further modeled by simple multiplicative model, involving the following assumptions. Under a naïve model cfDNAfragments are uniformly distributed by genomic position with stochastic sampling

process being the dominating factor contributing to coverage variation. Then, copy number signal, i.e. read count associated with a given probe can be modeled as:

[00108] Observed UMC = True UMC x Underlying positional cfDNA profile (bait, cfDNAfragment) x Ligation efficiency (position,size , cfDNAfragment) x pulldown efficiency (probe, cfDNAfragment) x PCR efficiency (DNA fragment) x flow cell clustering loss x demultiplexing loss and collapsing loss x alignment loss (cfDNAfragment sequence).

[00109] This model assumes a multiplicative nature of the above model. The underlying bait-specific copy number signal can be inferred from observed UMC (e.g., UMC of a given sequence capture probe) in relation to an established baseline by a series of steps, such as the baseline determination methods disclosed herein.

[00110] Methods disclosed herein provide approaches for estimating probe efficiency and bait saturation from the sample and training sets. Alternately, such parameters may be inferred by performing a set of bait titration experiments, where the effect of varying target sequence concentration on UMCs is observed for each probe. If K, I_{sat}, and UMC are known, it is then possible to determine a UMC value or range corresponding to tumor cells that have not undergone copy number variation. For example, under the assumption that most of the genetic loci have not undergone copy number alteration, the observed UMCs will largely be derived from diploid samples. Samples that have undergone copy number variation will be those genetic loci for which the UMCs fall outside the expected range for probes with their corresponding values of K and I_{sat}. In some cases, for example, the UMC value or range will be a function depending on K and I_{sat} for each probe. For example, the UMC corresponding to a diploid copy number can be different between two probes.

Computer control systems

[00111] The present disclosure provides computer control systems that are programmed to implement methods of the disclosure. **FIG. 12** shows a computer system 1201 that is programmed or otherwise configured to implement methods of the present disclosure. The computer system 1201 includes a central processing unit (CPU, also "processor" and "computer processor" herein) 1205, which can be a single core or multi core processor, or a plurality of processors for parallel processing. The computer system 1201 also includes memory or memory location 1210 (e.g., random-access memory, read-only memory, flash memory), electronic storage unit 1215 (e.g., hard disk), communication interface 1220 (e.g., network adapter) for communicating with one or more other systems, and peripheral devices 1225, such as cache, other memory, data storage and/or electronic display adapters. The memory 1210, storage unit 1215, interface 1220 and peripheral devices 1225 are in communication with the CPU 1205

through a communication bus (solid lines), such as a motherboard. The storage unit 1215 can be a data storage unit (or data repository) for storing data. The computer system 1201 can be operatively coupled to a computer network ("network") 1230 with the aid of the communication interface 1220. The network 1230 can be the Internet, an internet and/or extranet, or an intranet and/or extranet that is in communication with the Internet. The network 1230 in some cases is a telecommunication and/or data network. The network 1230 can include a local area network. The network 1230 can include one or more computer servers, which can enable distributed computing, such as cloud computing. The network 1230, in some cases with the aid of the computer system 1201, can implement a peer-to-peer network, which may enable devices coupled to the computer system 1201 to behave as a client or a server.

[00112] The CPU 1205 can execute a sequence of machine-readable instructions, which can be embodied in a program or software. The instructions may be stored in a memory location, such as the memory 1210. The instructions can be directed to the CPU 1205, which can subsequently program or otherwise configure the CPU 1205 to implement methods of the present disclosure. Examples of operations performed by the CPU 1205 can include fetch, decode, execute, and writeback.

[00113] The CPU 1205 can be part of a circuit, such as an integrated circuit. One or more other components of the system 1201 can be included in the circuit. In some cases, the circuit is an application specific integrated circuit (ASIC).

[00114] The storage unit 1215 can store files, such as drivers, libraries and saved programs. The storage unit 1215 can store user data, e.g., user preferences and user programs. The computer system 1201 in some cases can include one or more additional data storage units that are external to the computer system 1201, such as located on a remote server that is in communication with the computer system 1201 through an intranet or the Internet.

[00115] The computer system 1201 can communicate with one or more remote computer systems through the network 1230. For instance, the computer system 1201 can communicate with a remote computer system of a user. Examples of remote computer systems include personal computers (e.g., portable PC), slate or tablet PC's (e.g., Apple® iPad, Samsung® Galaxy Tab), telephones, Smart phones (e.g., Apple® iPhone, Android-enabled device, Blackberry®), or personal digital assistants. The user can access the computer system 1201 via the network 1230.

[00116] Methods as described herein can be implemented by way of machine (e.g., computer processor) executable code stored on an electronic storage location of the computer system 1201, such as, for example, on the memory 1210 or electronic storage unit 1215. The machine executable or machine readable code can be provided in the form of software. During use, the

code can be executed by the processor 1205. In some cases, the code can be retrieved from the storage unit 1215 and stored on the memory 1210 for ready access by the processor 1205. In some situations, the electronic storage unit 1215 can be precluded, and machine-executable instructions are stored on memory 1210.

[00117] The code can be pre-compiled and configured for use with a machine having a processer adapted to execute the code, or can be compiled during runtime. The code can be supplied in a programming language that can be selected to enable the code to execute in a pre-compiled or as-compiled fashion.

[00118] Aspects of the systems and methods provided herein, such as the computer system 1201, can be embodied in programming. Various aspects of the technology may be thought of as "products" or "articles of manufacture" typically in the form of machine (or processor) executable code and/or associated data that is carried on or embodied in a type of machine readable medium. Machine-executable code can be stored on an electronic storage unit, such as memory (e.g., read-only memory, random-access memory, flash memory) or a hard disk. "Storage" type media can include any or all of the tangible memory of the computers, processors or the like, or associated modules thereof, such as various semiconductor memories, tape drives, disk drives and the like, which may provide non-transitory storage at any time for the software programming. All or portions of the software may at times be communicated through the Internet or various other telecommunication networks. Such communications, for example, may enable loading of the software from one computer or processor into another, for example, from a management server or host computer into the computer platform of an application server. Thus, another type of media that may bear the software elements includes optical, electrical and electromagnetic waves, such as used across physical interfaces between local devices, through wired and optical landline networks and over various air-links. The physical elements that carry such waves, such as wired or wireless links, optical links or the like, also may be considered as media bearing the software. As used herein, unless restricted to non-transitory, tangible "storage" media, terms such as computer or machine "readable medium" refer to any medium that participates in providing instructions to a processor for execution.

[00119] Hence, a machine readable medium, such as computer-executable code, may take many forms, including but not limited to, a tangible storage medium, a carrier wave medium or physical transmission medium. Non-volatile storage media include, for example, optical or magnetic disks, such as any of the storage devices in any computer(s) or the like, such as may be used to implement the databases, etc. shown in the drawings. Volatile storage media include dynamic memory, such as main memory of such a computer platform. Tangible transmission media include coaxial cables; copper wire and fiber optics, including the wires that comprise a

bus within a computer system. Carrier-wave transmission media may take the form of electric or electromagnetic signals, or acoustic or light waves such as those generated during radio frequency (RF) and infrared (IR) data communications. Common forms of computer-readable media therefore include for example: a floppy disk, a flexible disk, hard disk, magnetic tape, any other magnetic medium, a CD-ROM, DVD or DVD-ROM, any other optical medium, punch cards paper tape, any other physical storage medium with patterns of holes, a RAM, a ROM, a PROM and EPROM, a FLASH-EPROM, any other memory chip or cartridge, a carrier wave transporting data or instructions, cables or links transporting such a carrier wave, or any other medium from which a computer may read programming code and/or data. Many of these forms of computer readable media may be involved in carrying one or more sequences of one or more instructions to a processor for execution.

[00120] The computer system 1201 can include or be in communication with an electronic display 1235 that comprises a user interface (UI) 1240 for providing, for example, a report. Examples of UI's include, without limitation, a graphical user interface (GUI) and web-based user interface.

[00121] Methods and systems of the present disclosure can be implemented by way of one or more algorithms. An algorithm can be implemented by way of software upon execution by the central processing unit 1205.

Examples

Example 1

[00122] Examination of previously-generated copy number variation spike-in data revealed significant probe-to-probe signal variation, both in raw read counts and UMCs, as well as the probe/gene-level copy number signal response to underlying copy number changes. See **FIG. 2**. **FIG. 3** illustrates the inferred versus theoretical copy number of three genes (CCND1, CCND2, and ERBB2), demonstrating the non-linear response of normalized coverage to the amount of bait in the sample. . These results suggest bait depletion during pulldown, which was confirmed by following bait titration effect in neighboring probes within the same gene with sizable differences in unique molecule count (thereby observing faster unique molecule count saturation for probes with high initial UMC).

[00123] **FIG. 4A** illustrates that the UMCs associated with each probe has a non-linear response with respect to probe *p*. **FIG. 4B** illustrates that UMCs associated with each probe have a non-linear response with respect to probe GC content.

[00124] **FIG. 5** illustrates UMCs of probes without performing saturation or probe-efficiency correction. **FIG. 6** shows the same sample after saturation correction. **FIG. 7** shows

the same sample after probe efficiency correction. The variance within genomic positions is reduced at each stage, leading to a clearer picture of the underlying copy number variation of the tumor cells emerging. Genes in **FIG. 7** for which the median probe post-probe efficiency correction signal is above 1.2 are called as having undergone copy number variation. Differing levels of post-probe efficiency correction signal are likely due to tumor heterogeneity or secondary tumors.

[00125] **FIG. 9** shows the typical progression of baselining genetic loci probe signal noise-reduction after saturation correction and probe efficiency correction.

[00126] **FIG. 10A** illustrates a plot of probe efficiency of the reference sample(s) on the x-axis and the sample's post-saturation corrected signal from a subject without copy-number variation in tumor cells. The relationship is approximately linear. **FIG. 10B** illustrates a similar plot from a subject with copy number variation in tumor cells. The response is not as linear as **FIG. 10A**. Correcting by the predicted efficiency inferred by determining the relationship between the probe efficiency from the reference sample(s) and the post-saturation corrected UMCs of the baselining genetic loci (indicated in black) will reduce variation due to differing probe efficiencies in the genetic loci that have putatively undergone copy number amplification in tumor cells (dots in grey). **FIG. 11** illustrates an exemplary report of copy number variation from a patient sample based on post-saturation and probe-efficiency corrected UMCs and MAF-optimized baselining. Stars indicate points that are indicated to belong to genetic loci that have undergone copy number variation in the tumor cells of the subject.

Example 2

[00127] Cell-free DNA is obtained from a subject with cancer, a barcoded sequencing library is prepared, a panel of oncogenes is enriched by sequence capture with a probe set, and the barcoded sequencing library is sequenced. The sequencing reads are mapped to a reference genome and collapsed into families based on their barcode sequences and mapping position. For each genomic coordinate corresponding to a midpoint of a probe from the probe set, the number of read families spanning that midpoint is counted to obtain a per-probe UMC. A median per-probe UMC is determined for each gene. To perform "saturation equilibrium correction," the genes are grouped by their median per-probe GC content. Genes for which the median per-probe UMCs differs significantly from those genes with similar median per-probe GC content are removed.

[00128] For each probe, p and GC content are determined as described herein. The remaining genes from the previous step are used to perform a two-dimension second-degree polynomial surface fit of the median gene-level UMC to probe p and GC content. The function relating p and GC content to an expected UMC is used to determine expected per-probe UMCs.

Residuals are determined for the data set by dividing the observed per-probe UMCs by the expected per-probe UMCs. The residual UMCs of each probe are the transformed quantitative measures of sequencing coverage.

[00129] Genes are again grouped by their median per-probe GC content, and genes whose median per-probe residual UMCs are significantly different from genes with similar median per-probe GC content are removed. "Probe efficiency" correction is then performed by obtaining residual UMCs of reference sample(s) as described in the preceding paragraphs. The residual UMCs of each probe from the sample are then divided by the residual UMCs of each corresponding probe from the reference(s) to obtain post-probe efficiency corrected UMCs.

[00130] Similar to saturation equilibrium correction above, the remaining genes are used to perform a two-dimension second-degree polynomial surface fit of the post-probe efficiency corrected UMC to probe *p* and GC content. The function relating *p* and GC content to an expected post-probe efficiency corrected UMC is used to determine an expected per-probe post-probe efficiency corrected UMC. Residuals are determined for the data set by dividing the observed per-probe post-probe efficiency corrected UMC by the expected per-probe post-probe efficiency corrected UMC. The residual post-probe efficiency corrected UMC of each probe are the post-probe GC-corrected signal.

[00131] The remaining genes are grouped by their median per-probe GC content, and genes whose median post-probe GC-corrected signal differs significantly from those genes with similar median per-probe GC content are removed.

[00132] The process of the example is repeated, with the post-probe GC-corrected signal as the starting input instead of the initial UMC.

[00133] For each gene, the median of the post-probe GC corrected signal is used to summarize each gene. Genes whose median post-probe GC-corrected signal is significantly different than the other genes are considered as candidates for having undergone gene amplification or deletion in the tumor cells.

[00134] For each gene, germline heterozygous alleles are determined and the relative frequency of each allele is quantified. Genetic loci used for baselining are found to have an approximately 1:1 ratio of alleles, validating the selection of baselining genetic loci.

[00135] A Z-score is determined for each gene based on the gene-level median post-probe GC-corrected signals and estimated standard deviations from whole-genome normal diploid probe signals. Genes with Z-scores higher than a cut-off are reported as having undergone gene amplification in tumor cells.

Example 3

[00136] The methods described herein were validated by measuring ERBB2 copy number in a method of the present disclosure against a control method. The method of the present disclosure produced a linear response of observed copy number (CN) vs. theoretical copy number, with no observed false positive CNV results in a normal (healthy) cohort. See **FIG. 13**, which shows the inferred gene copy number vs. the theoretical copy number estimate, with solid dots representing an observed copy number of ~2 (a diploid sample), open dots representing detected amplification events and the thick horizontal dashed line marking the mean gene CN cutoff. See also **FIG. 14**, which depicts the data of **FIG. 13**, with the control data represented by squares. All CNVs followed the expected titration trend down to 2.15 copies. Moreover, the method of the present disclosure decreased observed "noise" in the data due to a reduction in variance, allowing a CNV to be easily distinguished as compared to the control method. See the far right side of **FIG. 15**; triangles represent the method of the disclosure, while X's represent the control method.

[00137] While preferred embodiments of the present invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. It is not intended that the invention be limited by the specific examples provided within the specification. While the invention has been described with reference to the aforementioned specification, the descriptions and illustrations of the embodiments herein are not meant to be construed in a limiting sense. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the invention. Furthermore, it shall be understood that all aspects of the invention are not limited to the specific depictions, configurations or relative proportions set forth herein which depend upon a variety of conditions and variables. It should be understood that various alternatives to the embodiments of the invention described herein may be employed in practicing the invention. It is therefore contemplated that the invention shall also cover any such alternatives, modifications, variations or equivalents. It is intended that the following claims define the scope of the invention and that methods and structures within the scope of these claims and their equivalents be covered thereby.

CLAIMS

WHAT IS CLAIMED IS:

1. A method, comprising:

(a) obtaining sequencing reads of deoxyribonucleic acid (DNA) molecules of a cell-free bodily fluid sample of a subject;

(b) generating from the sequence reads a first data set comprising, for each genetic locus in a plurality of genetic loci, a quantitative measure related to sequencing read coverage ("read coverage");

(c) correcting the first data set by performing saturation equilibrium correction and probe efficiency correction;

(d) determining a baseline read coverage for the first data set, wherein the baseline read coverage relates to saturation equilibrium and probe efficiency; and

(e) determining a copy number state for each genetic locus in the plurality of genetic loci relative to the baseline read coverage.

2. The method of claim 1, wherein the first data set comprises, for each genetic locus in a plurality of genetic loci, a quantitative measure related to guanine-cytosine content of the genetic locus ("GC content").

3. The method of claim 2, comprising prior to (c) removing from the first data set genetic loci that are high-variance genetic loci, wherein removing comprises:

(i) fitting a model relating the quantitative measures related to guanine-cytosine content and the quantitative measures of sequencing read coverage of the genetic loci; and

(ii) removing from the first data set at least 10% of the genetic loci, wherein removing the genetic loci comprises removing the at least 10% of the genetic loci that most differ from the model, thereby providing the first data set of baselining genetic loci.

4. The method of claim 3, comprising removing at least 45% of the genetic loci.

5. The method of claim 3, wherein performing saturation equilibrium correction comprises transforming the first set data set of baselining genetic loci into a saturation corrected data set by:

(i) determining for each genetic locus from the first data set of baselining genetic loci a quantitative measure related to the probability that a strand of DNA molecule from the sample derived from the genetic locus is represented within the sequencing reads;

(ii) determining a first transformation for the read coverage by relating the read coverage in the first data set of baselining genetic loci to both the GC content of the first data set of baselining genetic loci and the quantitative measure related to the probability that a strand of DNA derived from each locus in the first data set of baselining genetic loci is represented within the sequencing reads; and

(iii) applying the first transformation to the read coverage of each genetic locus from the first data set of baselining genetic loci to provide the saturation corrected data set, wherein the saturation corrected data set comprises a first set of transformed read coverages of the first data set of baselining genetic loci;

6. The method of claim 5, wherein determining the first transformation comprises (i) determining a measure related to central tendency of the read coverage of the first data set of baselining genetic loci; (ii) determining a function that fits the measure related to central tendency of the read coverage of the first data set of baselining genetic loci based on the GC content of the genetic locus and the quantitative measure related to the probability that a strand of DNA derived from the genetic locus is represented within the sequencing reads; and (iii) for each genetic locus of the first data set of baselining genetic loci, determining a difference between the read coverage predicted by the function and the read coverage, wherein the difference is the transformed read coverage.

7. The method of claim 6, wherein the function is a surface approximation.

8. The method of claim 7, wherein the surface approximation is a two-dimensional second degree polynomial.

9. The method of claim 5, wherein performing probe efficiency correction comprises transforming the saturation corrected data set into a probe efficiency corrected data set by:

(i) removing from the saturation corrected data set genetic loci that are high-variance genetic loci with respect to the first set of transformed read coverages, thereby providing a second data set of baselining genetic loci;

(ii) determining a second transformation for the first set of transformed read coverages related to the probe efficiency of the second data set of baselining genetic loci; and

(iii) transforming the first set of transformed read coverages of the second data set of baselining genetic loci with the second transformation, thereby providing the probe efficiency corrected data set, wherein the probe efficiency corrected data set comprises a second set of transformed read coverages of the second data set of baselining genetic loci.

10. The method of claim 9, wherein removing from the first data set genetic loci that are high-variance genetic loci comprises:

(i) fitting a model relating the GC content and the first set of transformed read coverages of the saturation corrected data set; and

(ii) removing from saturation corrected data set at least 10% of the genetic loci, wherein the removing the genetic loci comprises removing genetic loci that most differ from the model, thereby providing the second data set of baselining genetic loci.

11. The method of claim 10, comprising removing at least 45% of the genetic loci.

12. The method of claim 9, wherein the probe efficiency is determined by performing the saturation equilibrium correction on one or more reference samples, wherein the probe efficiency is the transformed read coverage obtained by performing the saturation equilibrium correction.

13. The method of claim 12, wherein the one or more reference samples are cell-free bodily fluid samples from a subject without cancer.

14. The method of claim 12, wherein the one or more reference samples are cell-free bodily fluid samples from a subject with cancer, wherein the corresponding genetic locus has not undergone copy number alteration.

15. The method of claim 12, wherein determining the second transformation comprises (i) fitting the probe efficiency determined for the genetic loci from the one or more reference samples to the first set of read coverages from the second data set of baselining genetic loci; (ii) dividing the transformed read coverages of each genetic locus of the second data set of baselining genetic loci by a predicted probe efficiency based on the fitting of (i).

16. The method of claim 5, further comprising:

(g) determining a third transformation for the second set of transformed read coverages by relating the transformed read coverages of the second data set of baselining genetic loci to both the GC content of the second data set of baselining genetic loci and the quantitative measure related to the probability that a strand of DNA derived from the each locus in the second data set of baselining genetic loci is represented within the sequencing reads;

(h) applying the third transformation to the second set of transformed read coverages to provide a fourth data set, wherein the fourth data set comprises a third set of transformed quantitative read coverages.

17. The method of claim 1, wherein the DNA of the cell-free bodily fluid sample is enriched for the set of genetic loci using one or more oligonucleotide probes that are complementary to at least a portion of the genetic locus from the set of genetic loci.

18. The method of claim 17, wherein the GC content of each genetic locus from the set of genetic loci is a measure related to central tendency of guanine-cytosine content of the one or more oligonucleotide probes that are complementary to at least a portion of the genetic locus from the set of genetic loci.

19. The method of claim 17, wherein the read coverage of the genetic locus is a measure related to central tendency of the read coverage of regions of the genetic locus corresponding to the one or more oligonucleotide probes.

20. The method of claim 17, wherein the performing saturation equilibrium correction and the performing probe efficiency correction comprises fitting a Langmuir model, wherein the Langmuir model comprises probe efficiency (K) and saturation equilibrium constant (I_{sat}).

21. The method of claim 20, wherein K and I_{sat} are determined empirically for each oligonucleotide probe in the one or more oligonucleotide probes.

22. The method of claim 21, wherein performing saturation equilibrium correction and performing probe correction comprises fitting the read coverages of the genetic loci to the Langmuir model assuming that the genetic loci are present in identical copy number states, thereby providing a baseline read coverage.

23. The method of claim 22, wherein the identical copy number states are diploid.

24. The method of claim 22, wherein the baseline read coverage is a function dependent on the probe efficiency and the saturation equilibrium.

25. The method of claim 22, wherein determining a copy number state comprises comparing the read coverage of the genetic loci to the baseline read coverage.

26. The method of any one of the preceding claims, wherein the cell-free bodily fluid is selected from the group consisting of serum, plasma, urine, and cerebrospinal fluid.

27. The method of any one of the preceding claims, wherein the read coverage is determined by mapping the sequencing reads to a reference genome.

28. The method of any one of the preceding claims, wherein obtaining the sequencing reads comprises ligating adaptors to the DNA molecules from the cell-free bodily fluid from the subject.

29. The method of claim 28, wherein the DNA molecules are duplex DNA molecules and the adaptors are ligated to the duplex DNA molecules such that each adaptor differently tags complementary strands of the DNA molecule to provide tagged strands.

30. The method of claim 29, wherein determining the quantitative measure related to the probability that a strand of DNA derived from the genetic locus is represented within the sequencing reads comprises sorting sequencing reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in said set, and (ii) each unpaired read represents a first tagged strand having no second differently tagged complementary strand derived from a double-stranded polynucleotide molecule represented among said sequence reads in said set of sequence reads.

31. The method of claim 30, further comprising determining quantitative measures of (i) said paired reads and (ii) said unpaired reads that map to each of one or more genetic loci to determine a quantitative measure related to total double-stranded DNA molecules in said sample that map to each of said one or more genetic loci based on said quantitative measure related to paired reads and unpaired reads mapping to each locus.

32. The method of claim 28, wherein the adaptors comprise barcode sequences.

33. The method of claim 32, wherein determining the read coverage comprises collapsing the sequencing reads based on position of the mapping of the sequencing reads to the reference genome and the barcode sequences.

34. The method of any one of the preceding claims, wherein the genetic loci comprise one or more oncogenes.

35. The method of any one of the preceding claims, further comprising determining that at least a subset of the baselining genetic loci have undergone copy number alteration in the tumor cells of the subject by determining relative quantities of variants within the baselining genetic loci for which the germline genome of the subject is heterozygous.

36. The method of claim 35, wherein the relative quantities of the variants are not approximately equal.

37. The method of claim 36, wherein the baselining genetic loci for which the relative quantities of the variants are not approximately equal are removed from the baselining genetic loci, thereby providing allelic-frequency corrected baselining genetic loci.

38. The method of claim 37, wherein the allelic-frequency corrected baselining genetic loci are used as the baselining loci in the methods of any one of the preceding claims.

39. A method comprising:

(a) receiving into memory sequencing reads of deoxyribonucleic acid (DNA) molecules of a cell-free bodily fluid sample of a subject;

(b) executing code with a computer processor to perform the following steps:

(i) generating from the sequence reads a first data set comprising for each genetic locus in a plurality of genetic loci a quantitative measure related to sequencing read coverage ("read coverage");

(ii) correcting the first data set by performing saturation equilibrium correction and probe efficiency correction;

(iii) determining a baseline read coverage for the first data set, wherein the baseline read coverage relates to saturation equilibrium and probe efficiency; and

(iv) determining a copy number state for each genetic locus in the plurality of genetic loci relative to the baseline read coverage.

40. A system comprising:

(a) a network;

(b) a database comprising computer memory configured to store nucleic acid sequence data which are connected to the network;

(c) a bioinformatics computer comprising a computer memory and one or more computer processors, which computer is connected to the network;

wherein the computer further comprises machine-executable code which, when executed by the one or more computer processors, copies said nucleic acid sequence data stored in the database, writes the copied data to memory in the bioinformatics computer and performs steps including:

(i) generating from the nucleic acid sequence data a first data set comprising for each genetic locus in a plurality of genetic loci a quantitative measure related to sequencing read coverage ("read coverage");

(ii) correcting the first data set by performing saturation equilibrium correction and probe efficiency correction;

(iii) determining a baseline read coverage for the first data set, wherein the baseline read coverage relates to saturation equilibrium and probe efficiency; and

(iv) determining a copy number state for each genetic locus in the plurality of genetic loci relative to the baseline read coverage.

41. The system of claim 41, wherein said database is connected to a nucleic acid sequencer.

Point Mutations (SNVs) (70 Genes)							Amplifications (CNVs) (16 Genes)		Fusions (6 Genes)	Indels (3 Genes)
AKT1	*ALK*	*APC*	*AR*	*ARAF*	*ARID1A*		*AR*	*BRAF*	*ALK*	*EGFR**
ATM	*BRAF*	*BRCA1*	*BRCA2*	*CCND1*	*CCND2*		*CCNE1*	*CDK4*	*FGFR2*	*ERBB2**
CCNE1	*CDH1*	*CDK4*	*CDK6*	*CDKN2A*	*CDKN2B*		*CDK6*	*EGFR*	*FGFR3*	*MET***
CTNNB1	*EGFR*	*ERBB2*	*ESR1*	*EZH2*	*FBXW7*		*ERBB2*	*FGFR1*	*NTRK1*	
FGFR1	*FGFR2*	*FGFR3*	*GATA3*	*GNA11*	*GNAQ*		*FGFR2*	*KIT*	*RET*	
GNAS	*HNF1A*	*HRAS*	*IDH1*	*IDH2*	*JAK2*		*KRAS*	*MET*	*ROS1*	
JAK3	*KIT*	*KRAS*	*MAP2K1*	*MAP2K2*	*MET*		*MYC*	*PDGFRA*		
MLH1	*MPL*	*MYC*	*NF1*	*NFE2L2*	*NOTCH1*		*PIK3CA*	*RAF1*		
NPM1	*NRAS*	*NTRK1*	*PDGFRA*	*PIK3CA*	*PTEN*					
PTPN11	*RAF1*	*RB1*	*RET*	*RHEB*	*RHOA*					
RIT1	*ROS1*	*SMAD4*	*SMO*	*SRC*	*STK11*					
TERT	*TP53*	*TSC1*	*VHL*							

*exons 19 & 20
*exons 19 & 20
**exon 14 skipping

FIG. 1

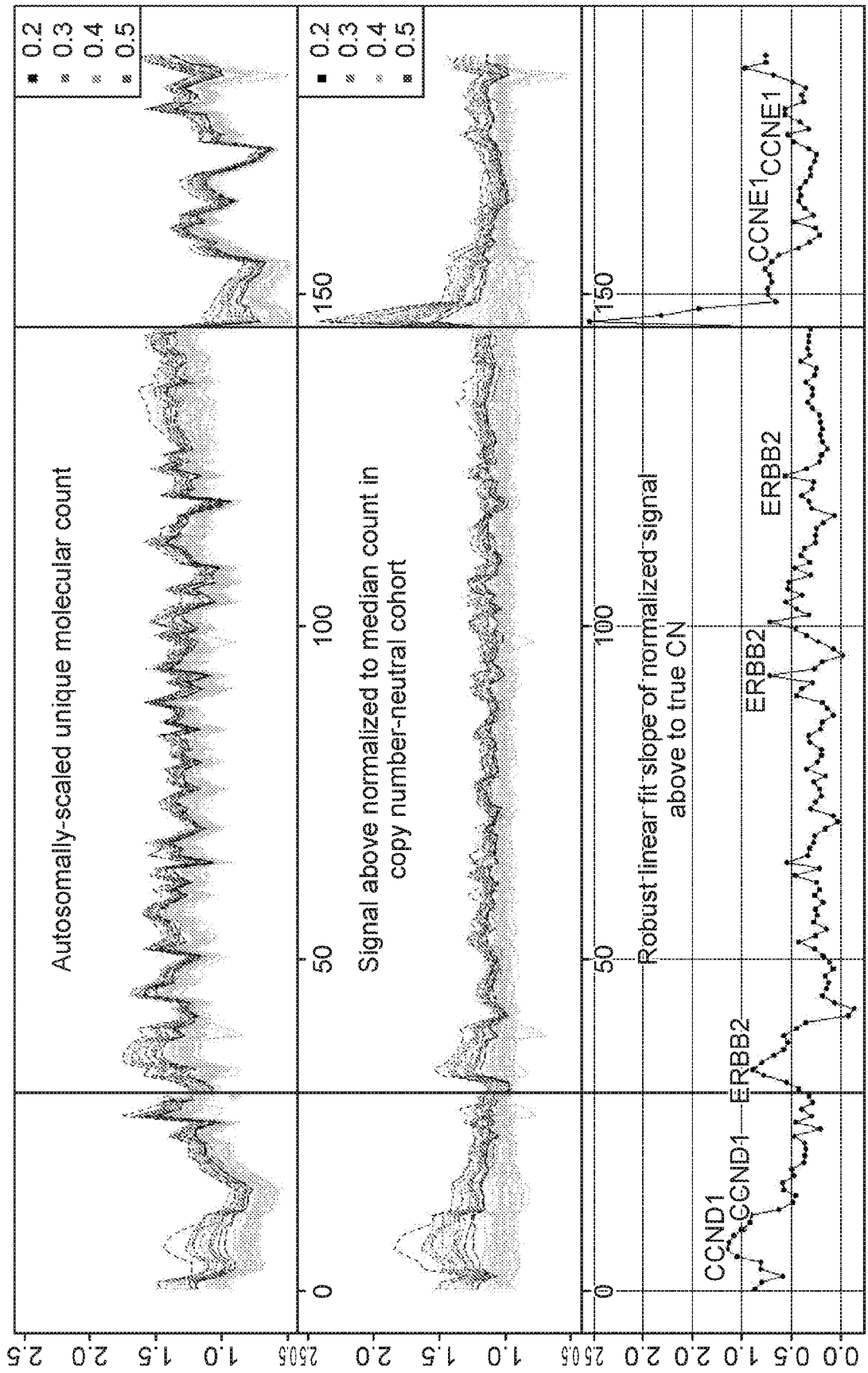
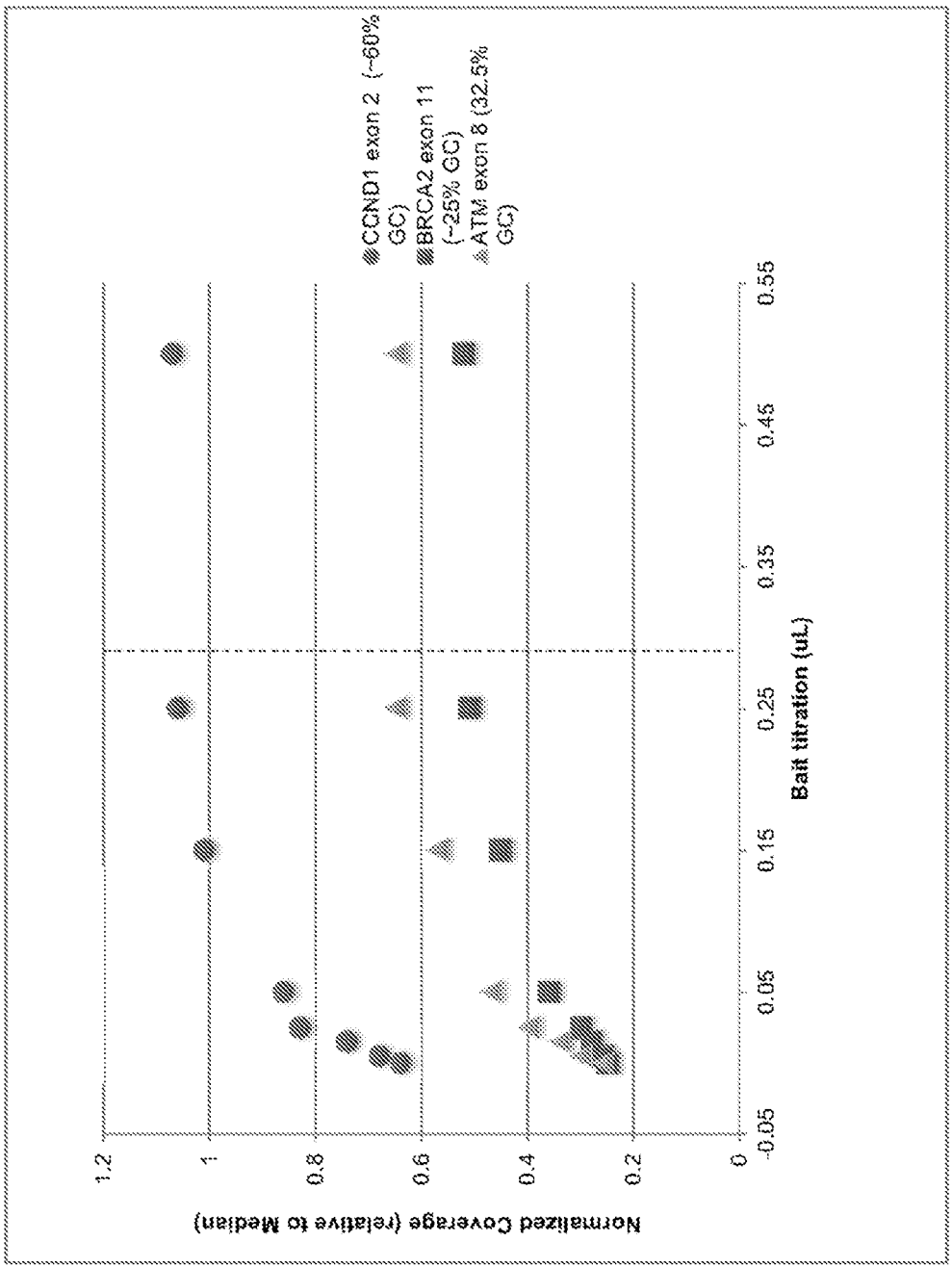

Autosomally-scaled unique molecular count
Signal above normalized to median count in copy number-neutral cohort
Robust linear fit slope of normalized signal above to true CN
0.2
0.3
0.4
0.5
CCND1
ERBB2
CCNE1
0
50
100
150
2505
0.0
0.5
1.0
1.5
2.0
2.5

FIG. 2

FIG. 3


CCND1 exon 2 (~60% GC)
BRCA2 exon 11 (~25% GC)
ATM exon 8 (32.5% GC)
Normalized Coverage (relative to Median)
1.2
1
0.8
0.6
0.4
0.2
0
-0.05
0.05
0.15
0.25
0.35
0.45
0.55
Bait titration (uL)

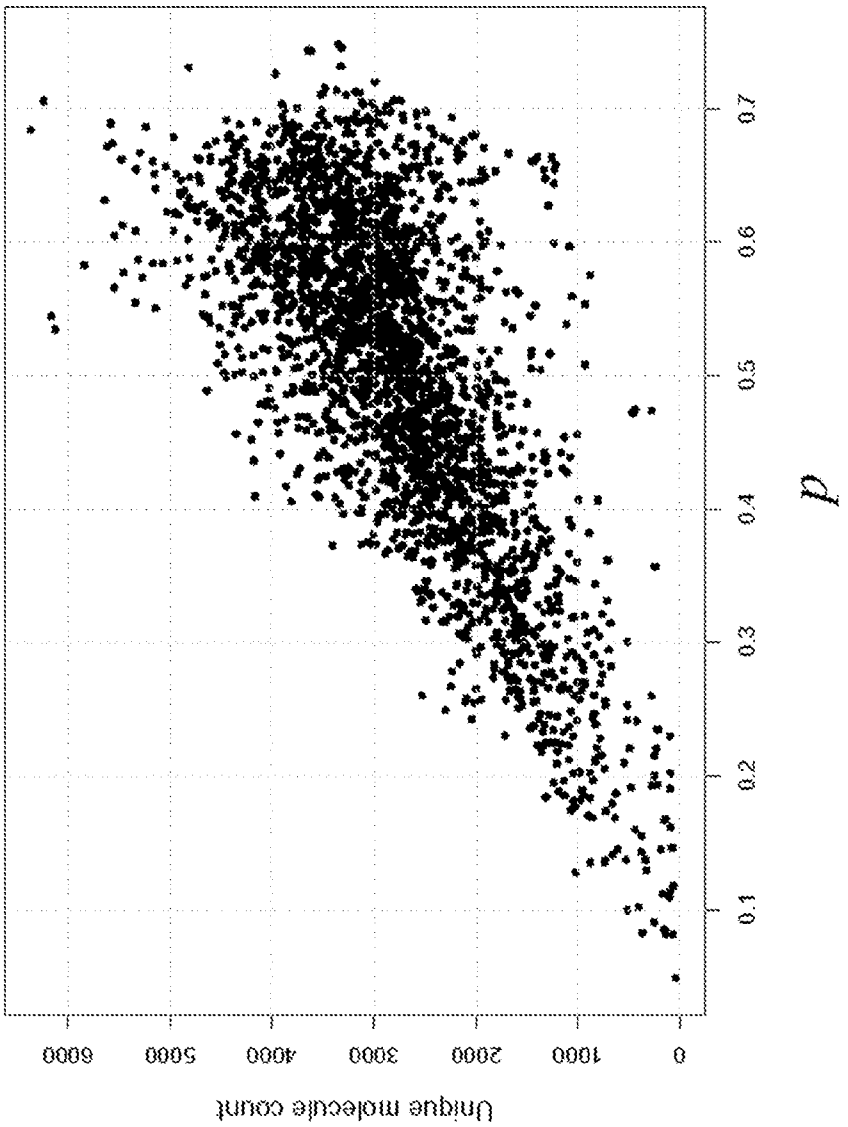

Unique molecule count
0
1000
2000
3000
4000
5000
6000
p
0.1
0.2
0.3
0.4
0.5
0.6
0.7

FIG. 4A

FIG. 4B

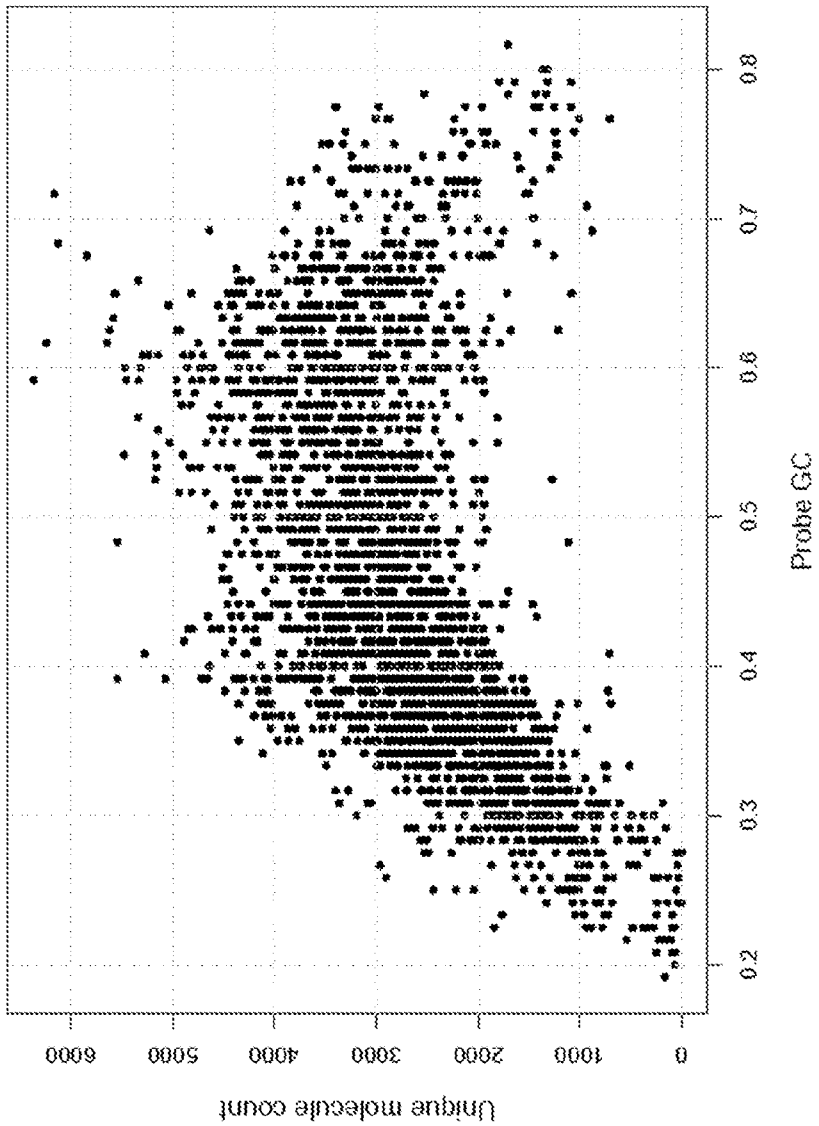

Probe GC
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Unique molecule count
0
1000
2000
3000
4000
5000
6000

FIG. 5

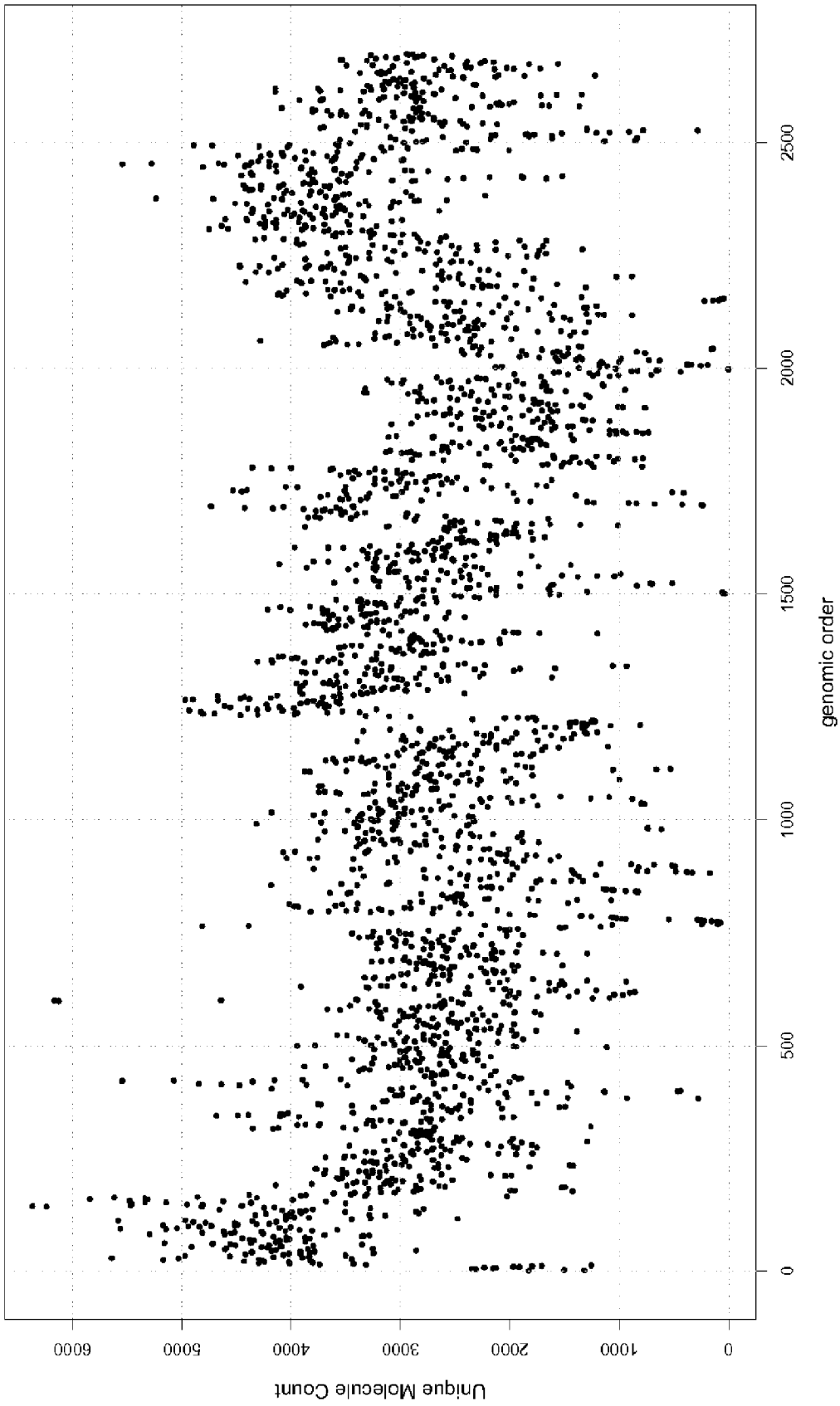

Unique Molecule Count
6000
5000
4000
3000
2000
1000
0
genomic order
0
500
1000
1500
2000
2500

FIG. 6

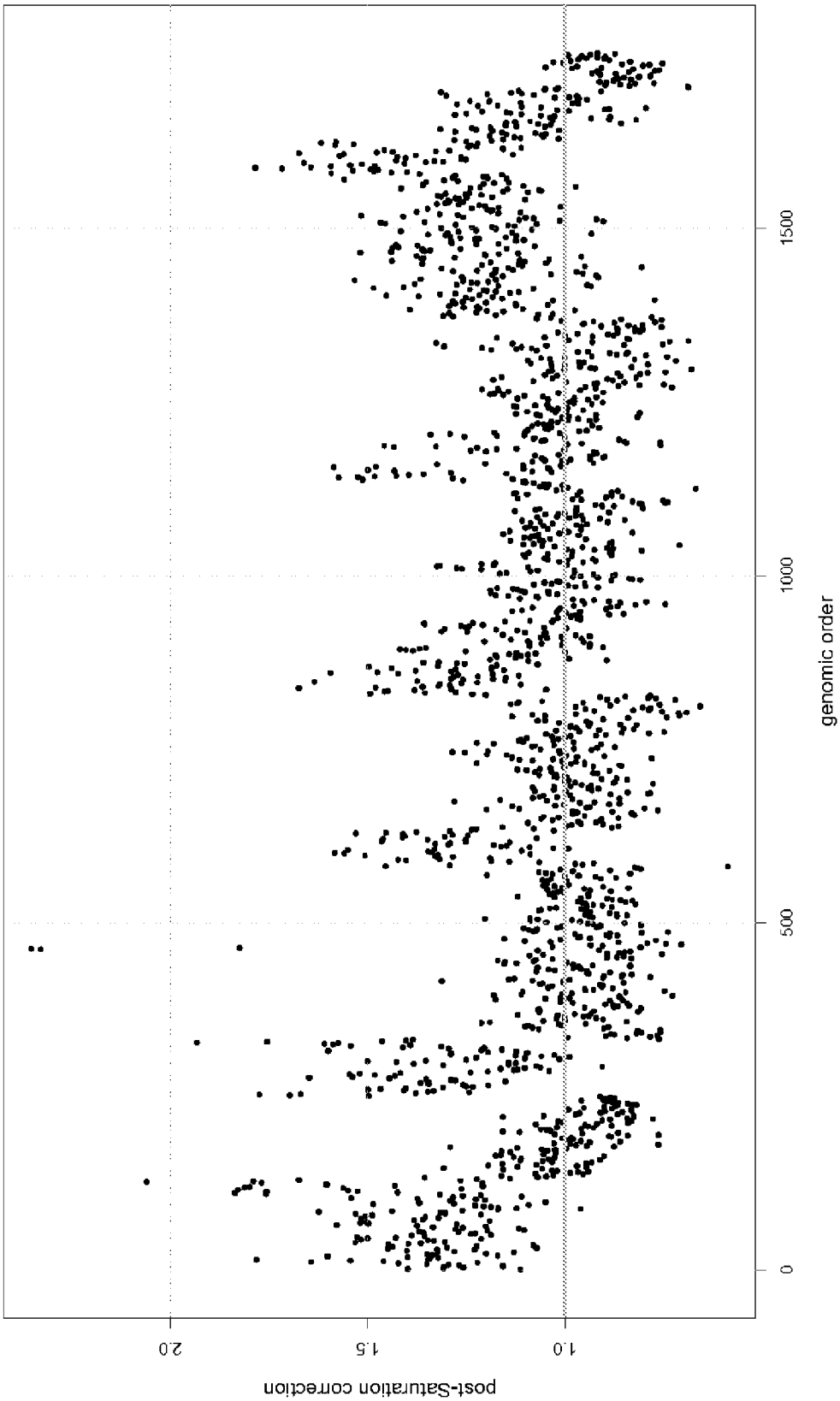
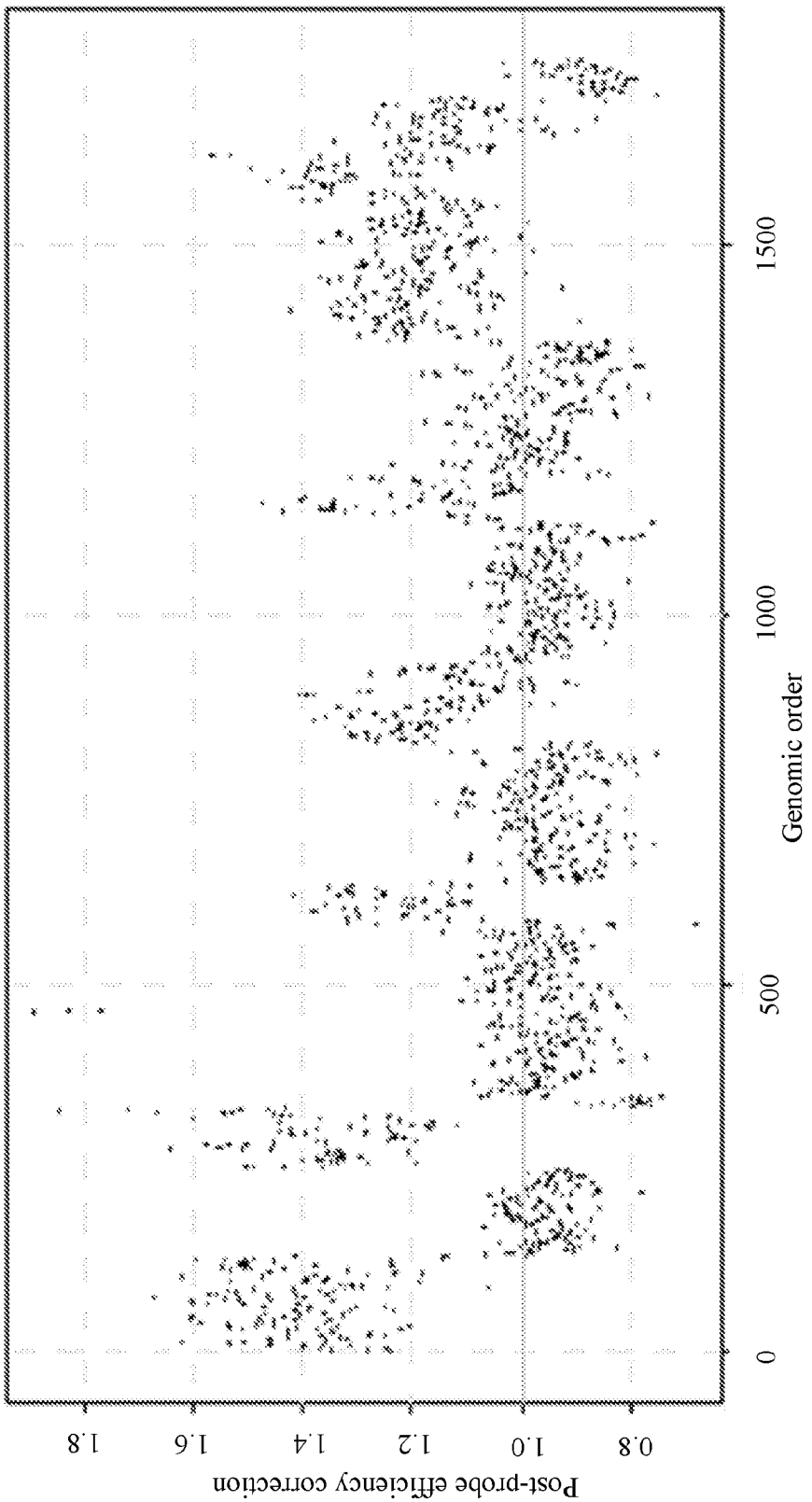

post-Saturation correction
1.0
1.5
2.0
genomic order
0
500
1000
1500

FIG. 7


Post-probe efficiency correction
0.8
1.0
1.2
1.4
1.6
1.8
Genomic order
0
500
1000
1500

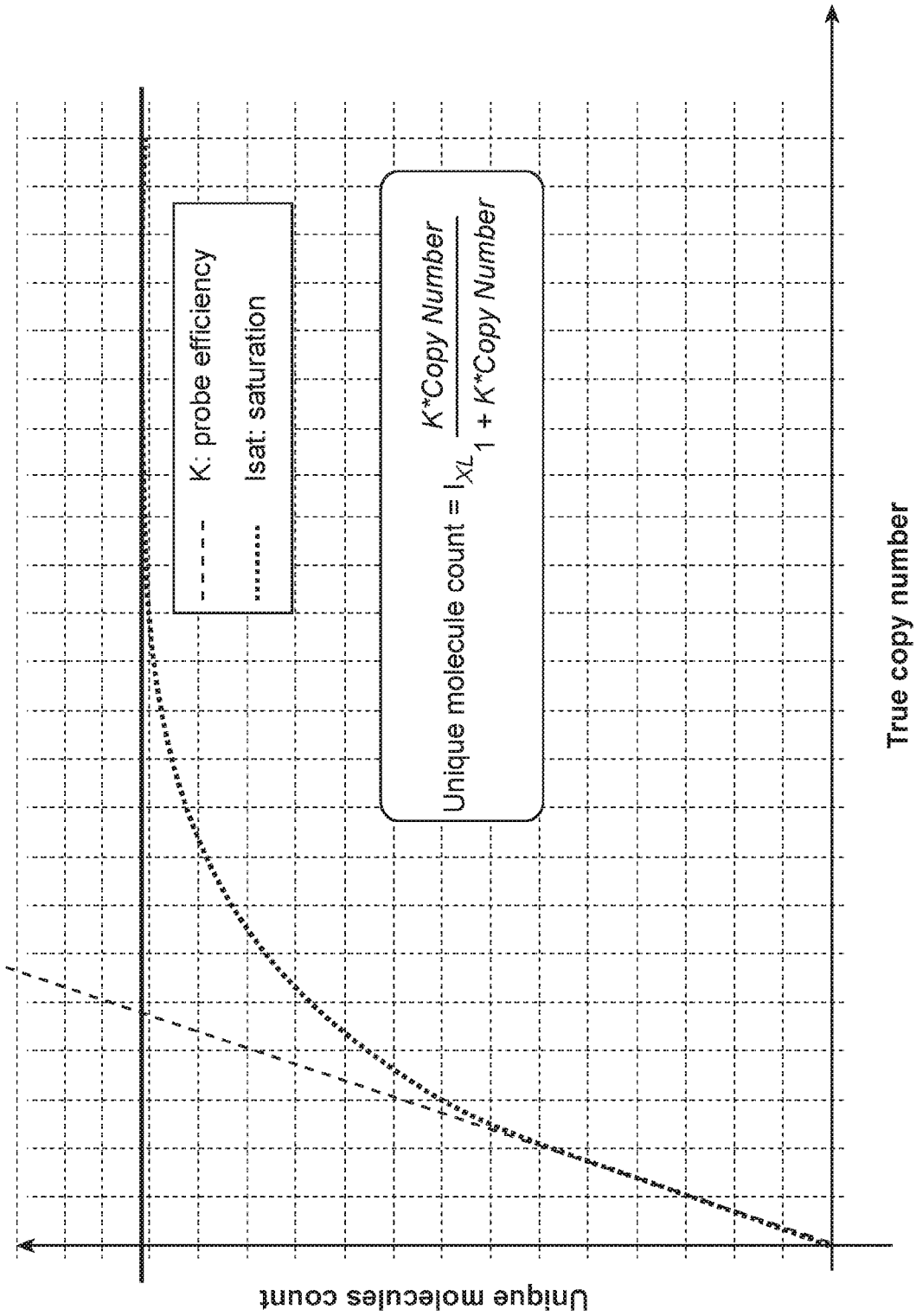

Unique molecules count
True copy number
K: probe efficiency
Isat: saturation
Unique molecule count = $I_{XL} \frac{K*Copy\ Number}{1 + K*Copy\ Number}$

FIG. 8

FIG. 9

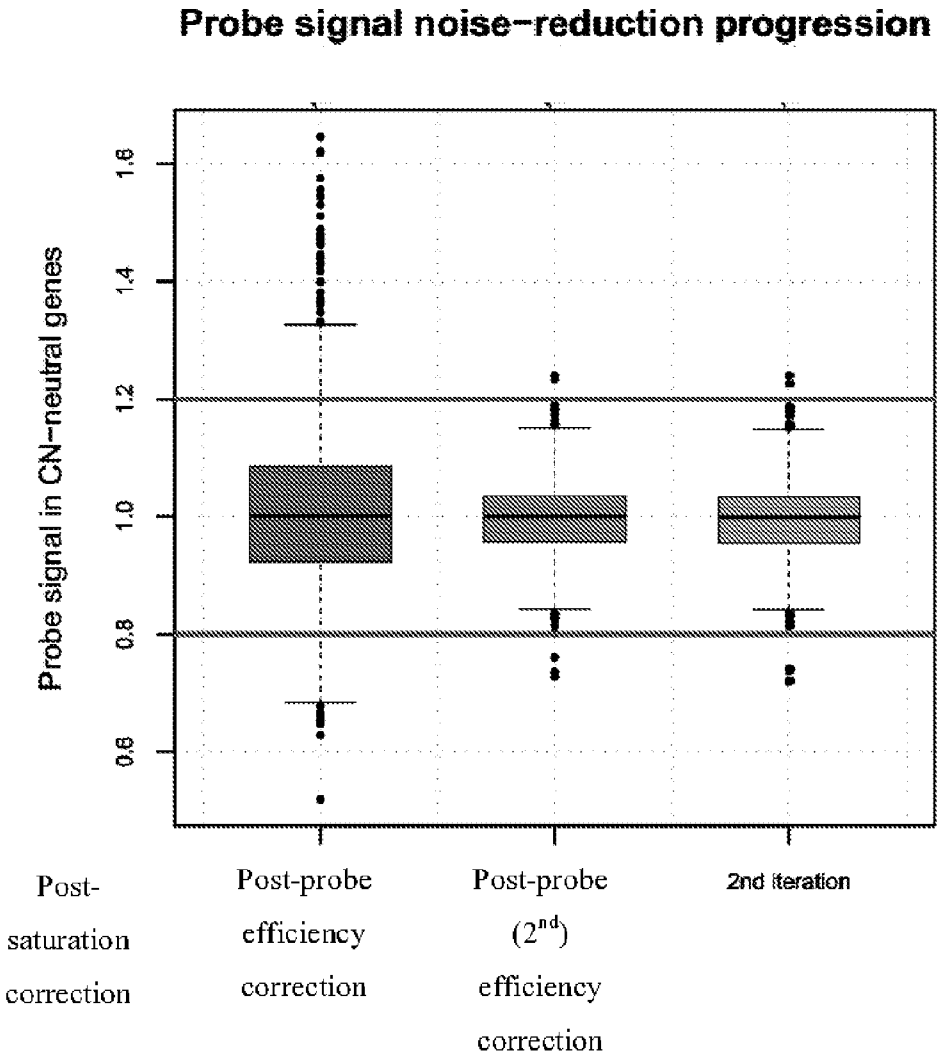

Probe signal noise-reduction progression
Probe signal in CN-neutral genes
0.6
0.8
1.0
1.2
1.4
1.6
Post-saturation correction
Post-probe efficiency correction
Post-probe (2nd) efficiency correction
2nd iteration

FIG. 10A

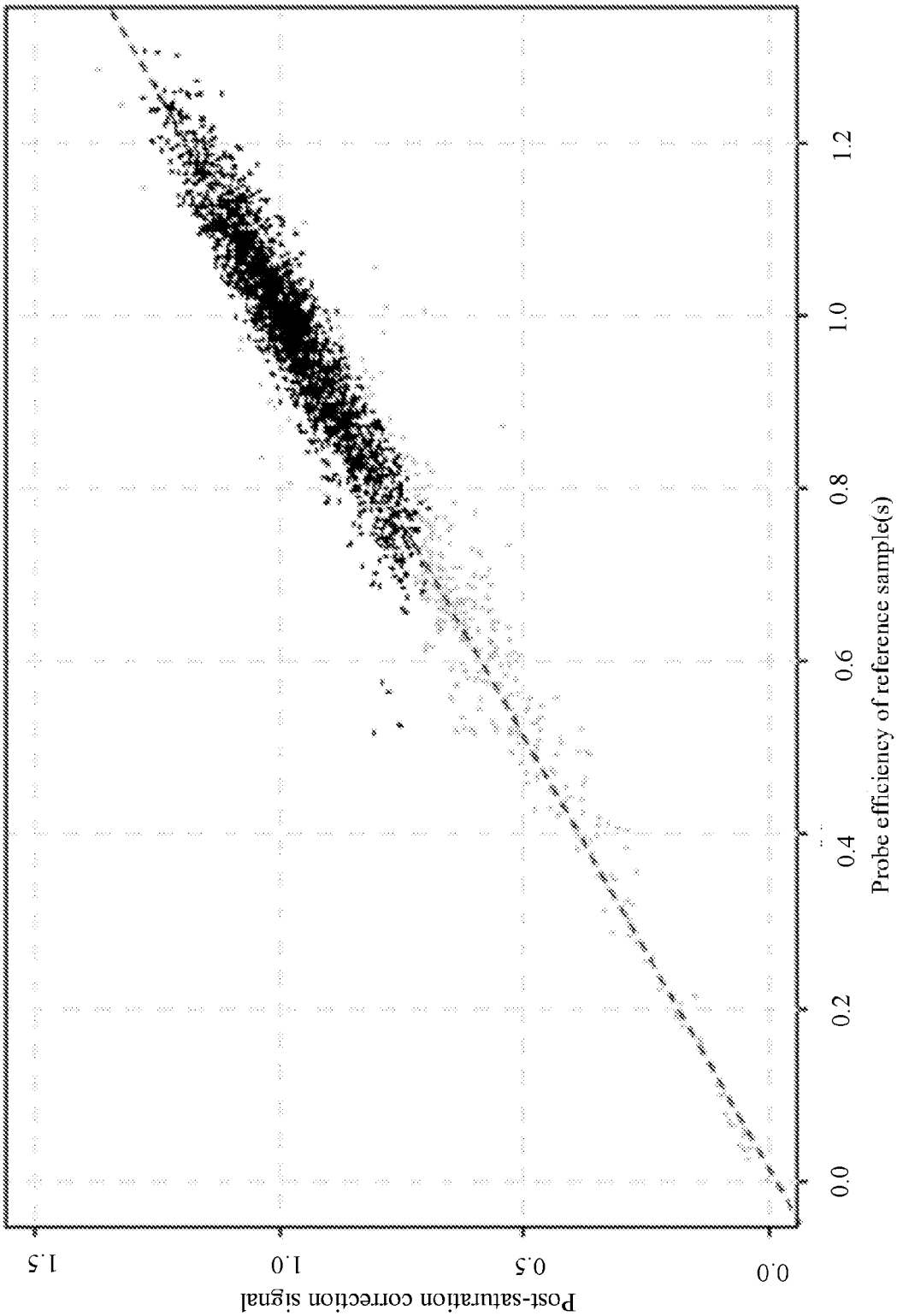
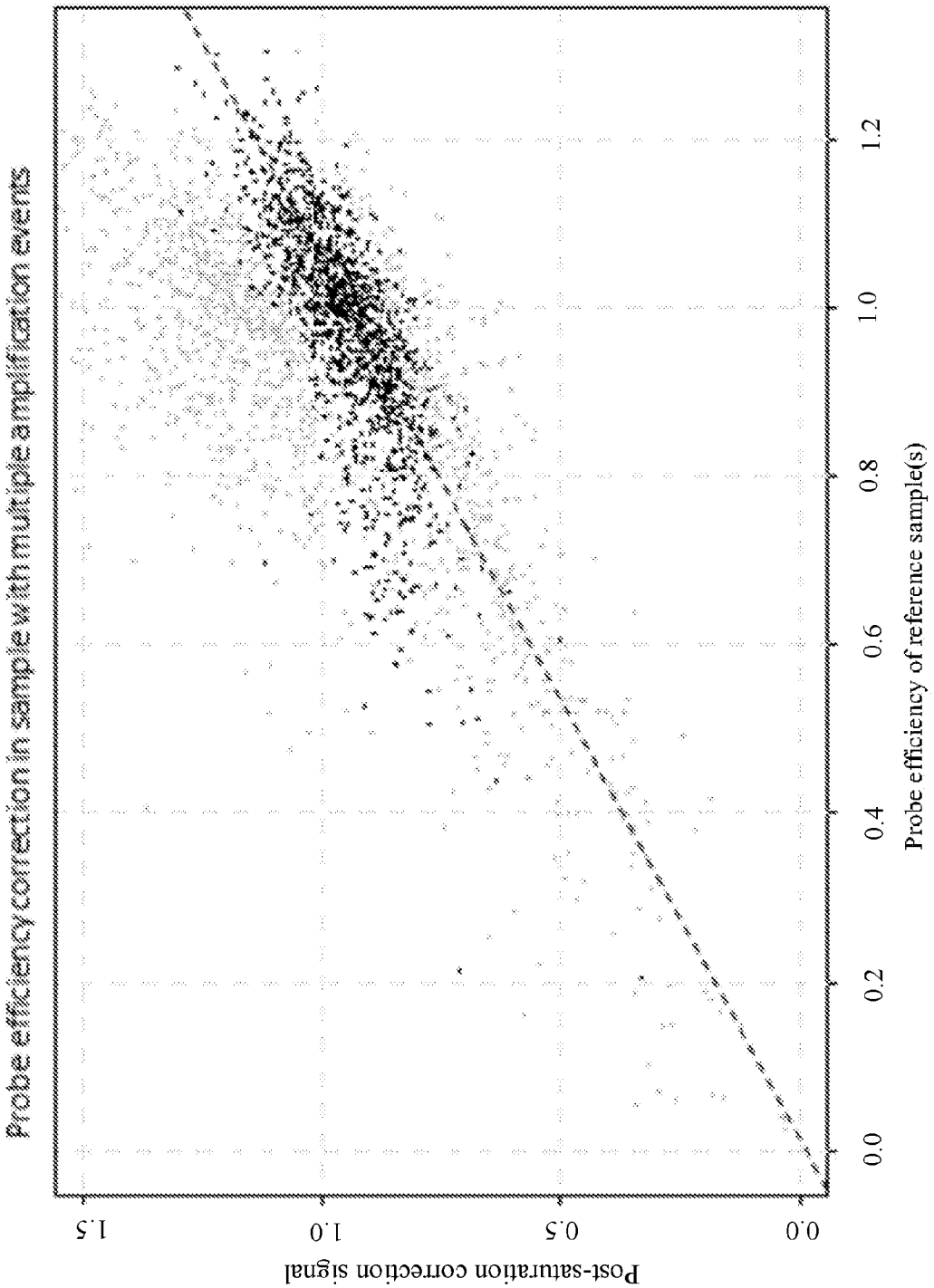

Post-saturation correction signal
0.0
0.5
1.0
1.5
Probe efficiency of reference sample(s)
0.0
0.2
0.4
0.6
0.8
1.0
1.2

FIG. 10B


Probe efficiency correction in sample with multiple amplification events
Post-saturation correction signal
0.0
0.5
1.0
1.5
Probe efficiency of reference sample(s)
0.0
0.2
0.4
0.6
0.8
1.0
1.2

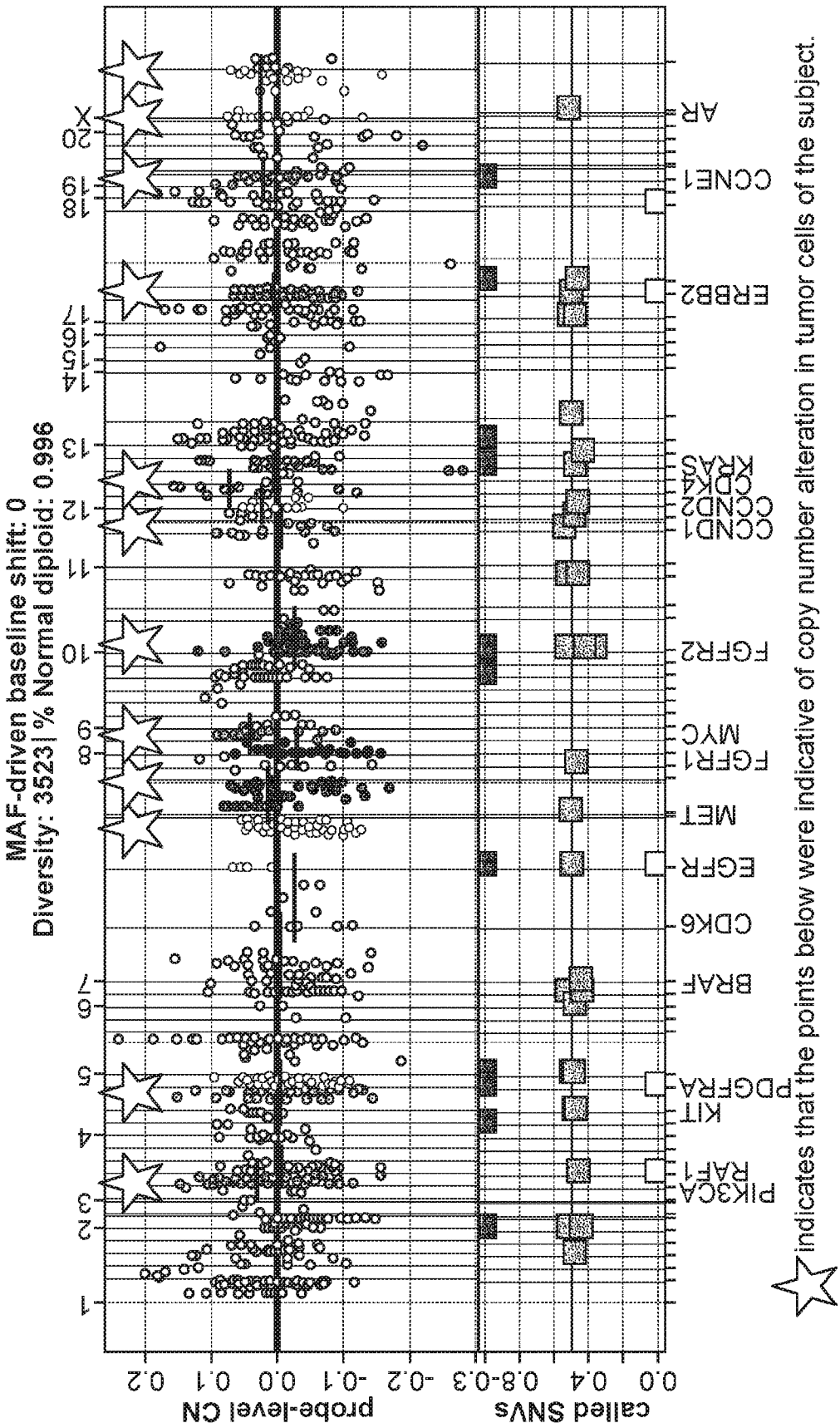

MAF-driven baseline shift: 0
Diversity: 3523| % Normal diploid: 0.996
probe-level CN
called SNVs
indicates that the points below were indicative of copy number alteration in tumor cells of the subject.
AR
CCNE1
ERBB2
CCND1
CCND2
CDK4
KRAS
FGFR2
FGFR1
MYC
MET
EGFR
CDK6
BRAF
PDGFRA
KIT
RAF1
PIK3CA

FIG. 11

FIG. 12

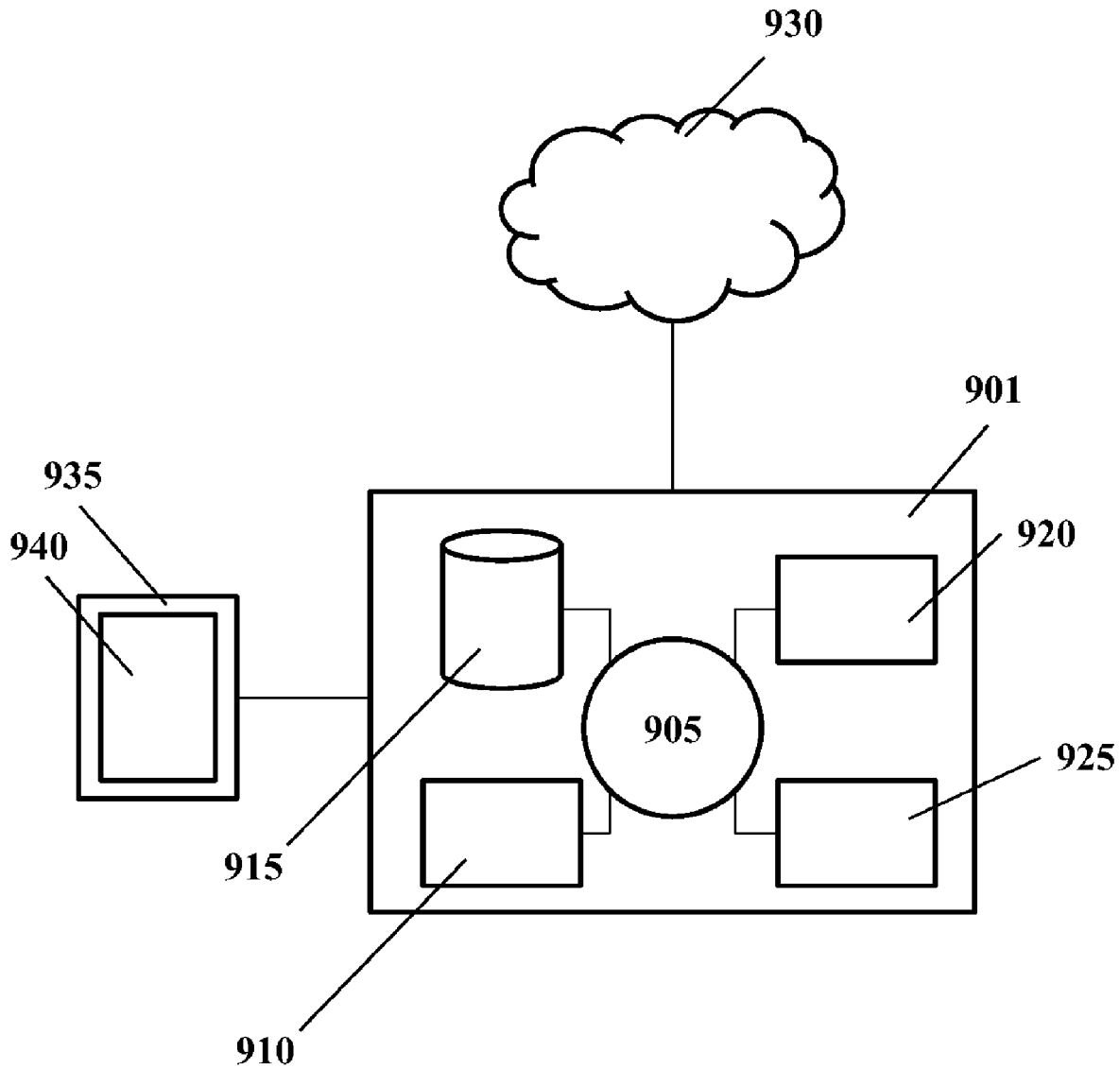


FIG. 13

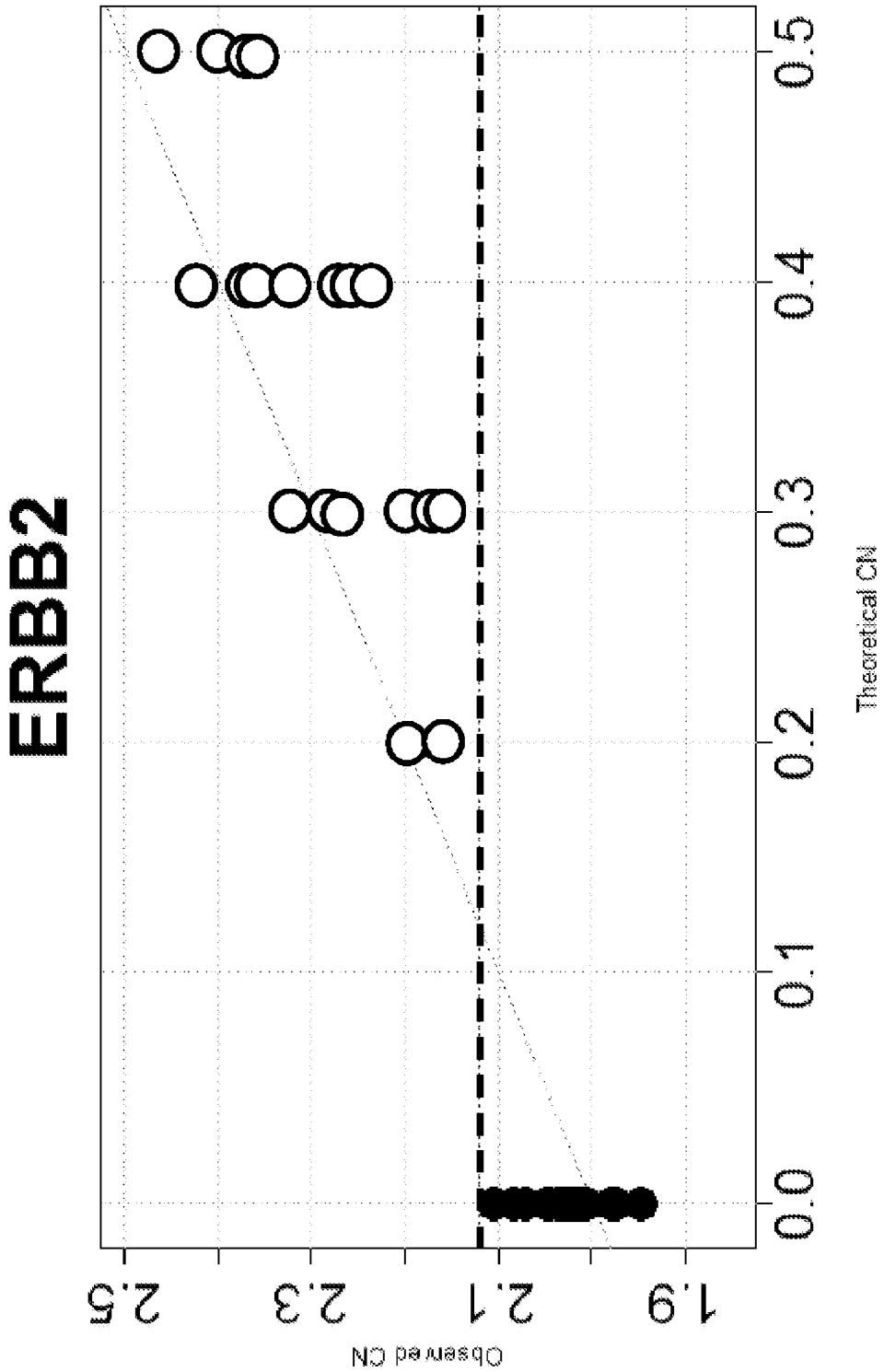
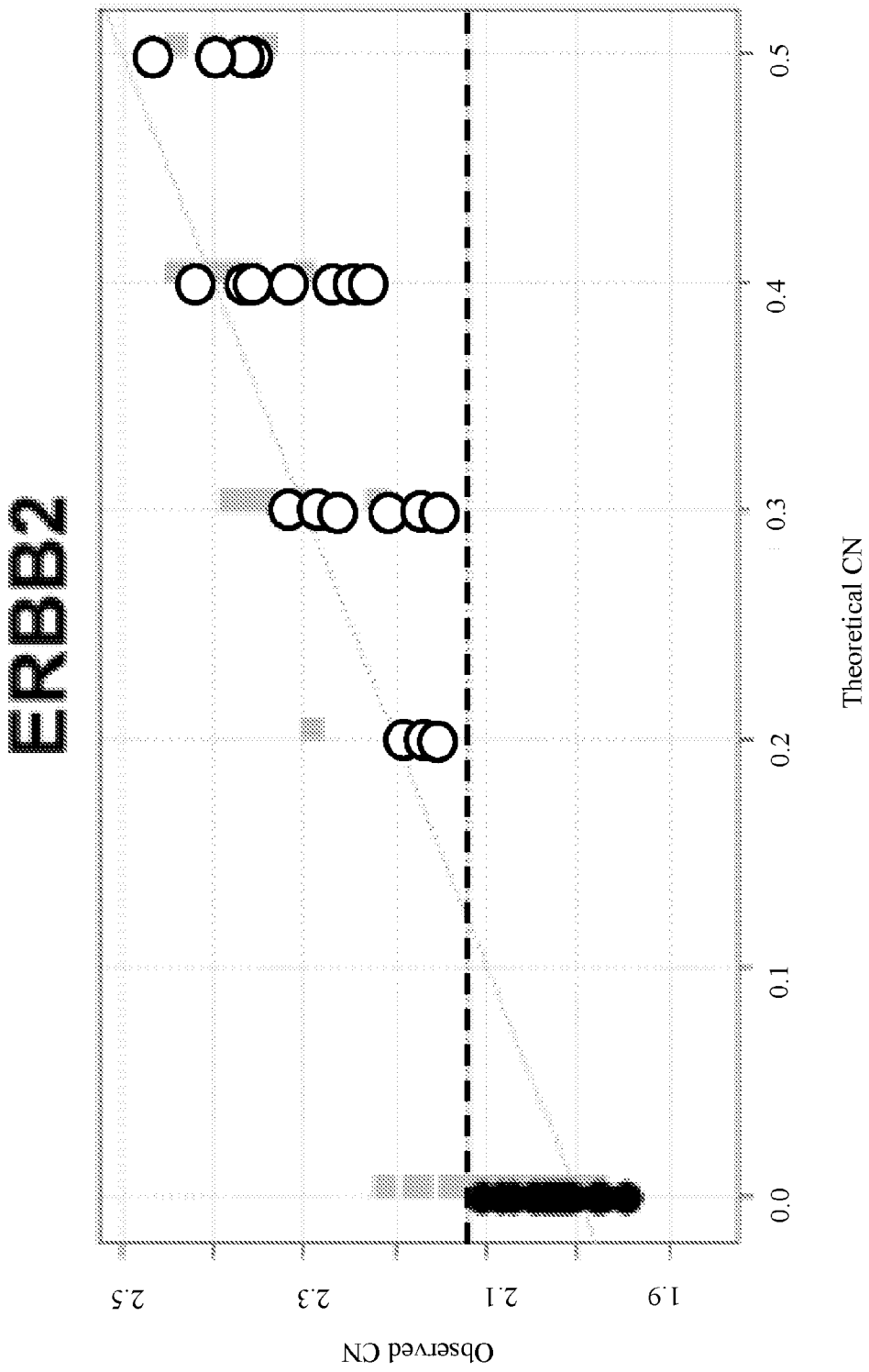


FIG. 14



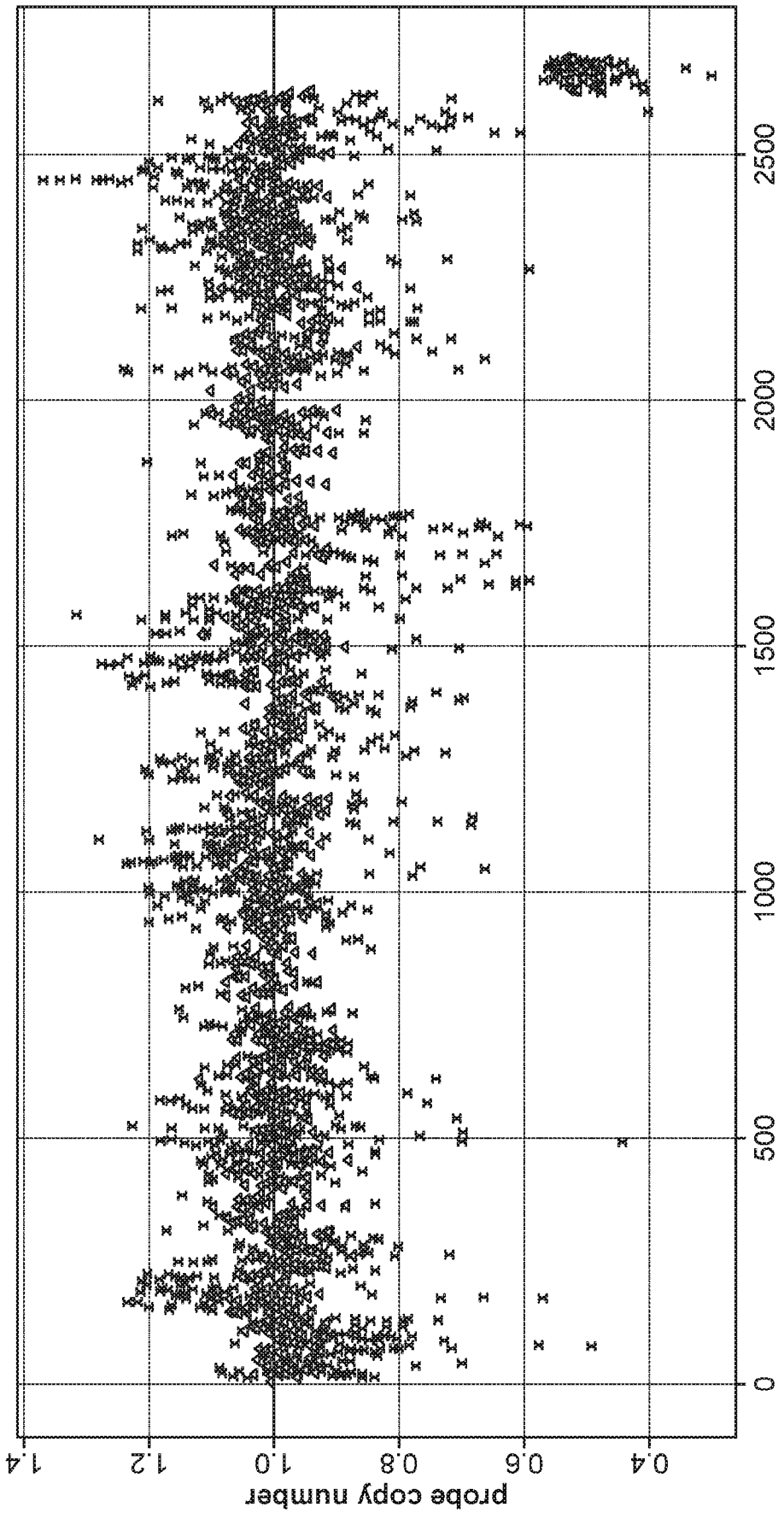

probe copy number
0.4
0.6
0.8
1.0
1.2
1.4
probes
0
500
1000
1500
2000
2500

FIG. 15

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 16/67356

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - C12Q 1/68, C12M 1/34, G06F 19/10, G01N 33/48 (2017.01)
CPC - C12Q 2537/16, G06F 19/10, C12Q 2537/165
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC(8) - C12Q 1/68, C12M 1/34, G06F 19/10, G01N 33/48 (2017.01)
CPC - C12Q 2537/16, G06F 19/10, C12Q 2537/165

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
CPC - C12Q 1/68, C12Q 2537/143, C12Q 2600/112, C12Q 1/6886
(keyword limited; terms below)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PatBase, Google Patents, Google Scholar
Search terms: copy number variation, copy number, sequence, sequencing, read coverage, saturation equlibrium, probe efficiency, Langmuir, model, algorith, correct, transform, normalize, computer, network, cell-free, sample, deoxyribonucleic acid, DNA, surface appro

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	US 2015/0299812 A1 (Guardant Health, Inc.) 22 October 2015 (22.10.2015) para [0004], [0008], [0009], [0021], [0033], [0056], [0068], [0105], [0109], [0115], [0118], [0149], [0155], [0201], [0203], [0204], [0209], [0210], [0211], [0269], [0272], [0274], [0276], [0277], [0288], [0300], [0323], [0324], [0328], [0331]	1-19, 26/(1-19), 39-41 ---------- 20-25, 26/(20-25)
Y	POZHITKOV et al., A Revised Design for Microarray Experiments to Account for Experimental Noise and Uncertainty of Probe Response. PLoS One. 2014, Vol 9, No 3, page e91295 (pp 1-10). Especially abstract; p 3, col 1, para 2; p 3, col 2, para 4	20-25, 26/(20-25)

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
20 February 2017	08 MAR 2017
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (January 2015)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 16/67356

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☒ Claims Nos.: 27-38
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.