



(51) International Patent Classification:

C12Q 1/6869 (2018.01)

Published:

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

(21) International Application Number:

PCT/US2021/030295

(22) International Filing Date:

30 April 2021 (30.04.2021)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/018,363

30 April 2020 (30.04.2020)

US

(71) Applicant: **GUARDANT HEALTH, INC.** [US/US]; 505

Penobscot Drive, Redwood City, California 94063 (US).

(72) Inventors: **ARTIERI, Carlo**; 505 Penobscot Drive, Redwood City, California 94063 (US). **KONG, Yu**; 505 Penobscot Drive, Redwood City, California 94063 (US). **JAI-MOVICH, Ariel**; 505 Penobscot Drive, Redwood City, California 94063 (US). **KURATA, Jessica**; 505 Penobscot Drive, Redwood City, California 94063 (US).

(74) Agent: **KANAKARAJ, Indhu** et al.; 505 Penobscot Drive, Redwood City, California 94063 (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, IT, JO, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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).

(54) Title: METHODS FOR SEQUENCE DETERMINATION USING PARTITIONED NUCLEIC ACIDS

(57) Abstract: DNA damage (e.g., cytosine deamination) can appear more frequently in hypermethylated partitions of DNA (e.g., cell-free DNA) samples, than in hypomethylated partitions. Embodiments include sequencing hypermethylated partitions and hypomethylated partitions wherein calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules from the hypermethylated partition requires observation of the transition mutation in a greater number of molecules than calling a C to T or G to A transition mutation relative to the reference sequence based on sequences of molecules from the hypomethylated partition, or C to T or G to A transition mutations are not called relative to a reference sequence based on sequences of molecules of the hypermethylated partition.



METHODS FOR SEQUENCE DETERMINATION USING PARTITIONED NUCLEIC ACIDS

[001] This application claims the benefit of US Provisional Patent Application No. 63/018,363, filed April 30, 2020, which is incorporated by reference in its entirety for all purposes.

BACKGROUND

[002] Cancer is responsible for millions of deaths per year worldwide. Early detection of cancer may result in improved outcomes because early-stage cancer tends to be more susceptible to treatment.

[003] Improperly controlled cell growth is a hallmark of cancer that generally results from an accumulation of genetic and epigenetic changes, such as copy number variations (CNVs), single nucleotide variations (SNVs), gene fusions, insertions and/or deletions (indels), epigenetic variations including DNA methylation such as 5-methylation of cytosine (5-methylcytosine) and association of DNA with chromatin proteins and transcription factors.

[004] Biopsies represent a traditional approach for detecting or diagnosing cancer in which cells or tissue are extracted from a possible site of cancer and analyzed for relevant phenotypic and/or genotypic features. Biopsies have the drawback of being invasive.

[005] Detection of cancer based on analysis of body fluids (“liquid biopsies”), such as blood, is an intriguing alternative based on the observation that DNA from cancer cells is released into body fluids. A liquid biopsy is noninvasive (perhaps requiring only a blood draw). However, it has been challenging to develop accurate and sensitive methods for analyzing liquid biopsy material given the low concentration and heterogeneity of cell-free DNA. This is especially true for procedures in which epigenetic changes including DNA methylation are analyzed by sequencing hypermethylated and hypomethylated partitions, as it has now been determined that DNA exhibiting a high degree of methylation can have damaged bases such as deaminated bases unrelated to the actual genomic sequence at a greater frequency than other DNA, which can adversely impact the accuracy of sequence determinations. Accordingly, there is a need for improved methods for sequence determinations using partitioned nucleic acids.

SUMMARY

[006] The present disclosure provides embodiments including methods of analyzing a sample of DNA, such as cell-free DNA (cfDNA), wherein the sample is partitioned into a plurality of partitions comprising a hypermethylated partition and a hypomethylated partition. The present disclosure is based in part on the following realization. In hypermethylated partitions, DNA (e.g.,

cfDNA) can have a higher amount of damage, such as cytosine deamination, that does not reflect an actual mutation in the cell from which the DNA originated. Such DNA damage can result in an increased frequency of apparent, but false-positive, C to T and complementary G to A transition mutations. Accordingly, it can be beneficial to use more stringent requirements for identifying such transition mutations based on sequences from a hypermethylated partition than for identifying such transition mutations based on sequences from a hypomethylated partition. Accordingly, the following embodiments are provided. Embodiment 1 is a method of analyzing a sample of DNA, the method comprising:

- partitioning the DNA sample into a plurality of partitions, wherein the plurality of partitions comprises a hypermethylated partition and a hypomethylated partition;

- tagging the DNA in the hypermethylated and hypomethylated partitions to generate tagged nucleic acids, wherein the tagged nucleic acids comprise molecular barcodes;

- obtaining sequence reads of molecules from the hypermethylated partition and sequence reads of molecules from the hypomethylated partition, wherein the sequence reads comprise molecular barcode sequence and sample sequence;

- grouping sequence reads into families based on at least one of (a) the molecular barcode sequences and (b) genomic positions corresponding to the first and last nucleotides of the sample sequence, wherein the families comprise sequence reads derived from a single DNA molecule in the sample;

- determining a first set of sequences of molecules from the hypermethylated partition and a second set of sequences of molecules from the hypomethylated partition; and

- calling a plurality of bases based on the first and second sets of sequences, wherein:

- (i) calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set requires observation of the transition mutation in a greater number of molecules than calling a C to T or G to A transition mutation relative to the reference sequence based on sequences of molecules of the second set; or

- (ii) C to T or G to A transition mutations are not called relative to a reference sequence based on sequences of molecules of the first set, or C to T or G to A transition mutations are called relative to a reference sequence based on sequences of molecules of the second set without the use of sequences of molecules of the first set, or a C to T or G to A

transition mutation is called relative to a reference sequence only if at least one sequence of a molecule of the second set comprises the C to T or G to A transition mutation.

[007] Embodiment 2 is the method of the immediately preceding embodiment, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set requires observation of the transition mutation in a greater number of molecules than calling a C to T or G to A transition mutation relative to the reference sequence based on sequences of molecules of the second set.

[008] Embodiment 3 is the method of any one of the preceding embodiments, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set requires observation of the transition mutation in at least three molecules.

[009] Embodiment 4 is the method of the immediately preceding embodiment, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set requires observation of the transition mutation in at least four molecules.

[0010] Embodiment 5 is the method of the immediately preceding embodiment, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set requires observation of the transition mutation in at least five molecules.

[0011] Embodiment 6 is the method of any one of the preceding embodiments, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the second set requires observation of the transition mutation in at least two molecules.

[0012] Embodiment 7 is the method of the immediately preceding embodiment, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the second set requires observation of the transition mutation in at least three molecules.

[0013] Embodiment 8 is the method of any one of the preceding embodiments, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set requires observation of the transition mutation in at least two more

molecules than does calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the second set.

[0014] Embodiment 9 is the method of any one of the preceding embodiments, wherein a first threshold is used for calling a C to T or G to A transition based on sequences of molecules of the first set and a second threshold is used for calling a C to T or G to A transition based on sequences of molecules of the second set; the first threshold provides a first level of specificity for calling a C to T or G to A transition; the second threshold provides a second level of specificity for calling a C to T or G to A transition; and the first level of specificity is approximately equal to the second level of specificity, or the first level of specificity is within 10%, 5%, 4%, 3%, 2%, 1%, 0.5%, 0.2%, or 0.1% of the second level of specificity.

[0015] Embodiment 10 is the method of the immediately preceding embodiment, wherein the first and second thresholds are specific for C to T and/or G to A transitions.

[0016] Embodiment 11 is the method of embodiment 9 or 10, wherein the first and second thresholds are determined from at least one control sample or a plurality of control samples, optionally wherein the at least one control sample or plurality of control samples are from individuals not suspected of having cancer.

[0017] Embodiment 12 is the method of any one of embodiments 1-8, wherein a first group of position-specific background error rates are used for a plurality of positions for sequences of molecules of the first set; a second group of position-specific background error rates are used for a plurality of positions for sequences of molecules of the second set; the second group comprises position-specific background error rates higher than the corresponding position-specific background error rates of the first group; and calling a C to T or G to A transition mutation based on sequences of molecules of the first set requires observation of the C to T or G to A transition mutation at a frequency exceeding the corresponding rate from the first group of position-specific background error rates.

[0018] Embodiment 13 is the method of the immediately preceding embodiment, wherein calling a C to T or G to A transition mutation based on sequences of molecules of the first set requires observation of the C to T or G to A transition mutation at a frequency exceeding the corresponding rate from the first group of position-specific background error rates by a factor of at least 2, 3, 4, or 5.

- [0019] Embodiment 14 is the method of the immediately preceding embodiment, wherein calling a C to T or G to A transition mutation based on sequences of molecules of the first set requires observation of the C to T or G to A transition mutation at a frequency exceeding the corresponding rate from the first group of position-specific background error rates by an amount consistent with a confidence level of at least 95%, 98%, 99%, 99.5%, or 99.9%.
- [0020] Embodiment 15 is the method of any one of embodiments 12-14, wherein the first and second groups of position-specific background error rates are determined from a plurality of control samples, optionally wherein the control samples are from individuals not suspected of having cancer.
- [0021] Embodiment 16 is the method of any one of embodiments 12-14, wherein the first and second groups of position-specific background error rates were determined using a plurality of control samples, optionally wherein the control samples are from individuals not suspected of having cancer.
- [0022] Embodiment 17 is the method of any one of embodiments 12-14, wherein the first and second groups of position-specific background error rates were determined using historical data.
- [0023] Embodiment 18 is the method of any one of embodiments 12-14, wherein the first and second groups of position-specific background error rates were determined using reads and/or sequences of molecules from the hypermethylated and hypomethylated partitions, respectively.
- [0024] Embodiment 19 is the method of any one of the preceding embodiments, further comprising obtaining sequence reads of molecules from an intermediate partition; determining a third set of sequences of molecules from the intermediate partition; and calling a plurality of bases based on the third set of sequences.
- [0025] Embodiment 20 is the method of the immediately preceding embodiment, wherein C to T and G to A transition mutations are called based on sequences of the third set less stringently than C to T and G to A transition mutations are called based on sequences of molecules of the first set.
- [0026] Embodiment 21 is the method of the immediately preceding embodiment, wherein C to T and G to A transition mutations are called based on sequences of the third set in the same way as C to T and G to A transition mutations are called based on sequences of the second set or more stringently than C to T and G to A transition mutations are called based on sequences of the second set.

[0027] Embodiment 22 is a method of analyzing a sample of DNA, the method comprising:
obtaining first and second sets of sequence reads from hypermethylated and
hypomethylated partitions of the sample, respectively; and

determining a sequence from the first and second sets of sequence reads, wherein:

(i) calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set requires observation of the transition mutation in a greater number of reads than calling a C to T or G to A transition mutation relative to the reference sequence based on reads of the second set; or

(ii) C to T or G to A transition mutations are not called relative to a reference sequence based on reads of the first set, or C to T or G to A transition mutations are called relative to a reference sequence based on sequences of molecules of the second set without the use of sequences of molecules of the first set, or a C to T or G to A transition mutation is called relative to a reference sequence only if at least one sequence of a molecule of the second set comprises the C to T or G to A transition mutation.

[0028] Embodiment 23 is the method of the immediately preceding embodiment, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set requires observation of the transition mutation in a greater number of reads than calling a C to T or G to A transition mutation relative to the reference sequence based on reads of the second set.

[0029] Embodiment 24 is the method of any one of embodiments 22 or 23, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set requires observation of the transition mutation in at least three reads.

[0030] Embodiment 25 is the method of the immediately preceding embodiment, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set requires observation of the transition mutation in at least four reads.

[0031] Embodiment 26 is the method of the immediately preceding embodiment, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set requires observation of the transition mutation in at least five reads.

[0032] Embodiment 27 is the method of any one of embodiments 22-26, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the second set requires observation of the transition mutation in at least two reads.

- [0033] Embodiment 28 is the method of the immediately preceding embodiment, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the second set requires observation of the transition mutation in at least three reads.
- [0034] Embodiment 29 is the method of any one of embodiments 22-28, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set requires observation of the transition mutation in at least two more reads than does calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the second set.
- [0035] Embodiment 30 is the method of any one of the preceding embodiments, further comprising obtaining a third set of sequence reads from an intermediate partition, wherein the sequence is determined from the third set in addition to the first and second sets.
- [0036] Embodiment 31 is the method of the immediately preceding embodiment, wherein C to T and G to A transition mutations are called based on reads of the third set less stringently than C to T and G to A transition mutations are called based on reads of the first set.
- [0037] Embodiment 32 is the method of the immediately preceding embodiment, wherein C to T and G to A transition mutations are called based on reads of the third set in the same way as C to T and G to A transition mutations are called based on reads of the second set.
- [0038] Embodiment 33 is the method of any one of the preceding embodiments, wherein the DNA of the hypermethylated partition and the DNA of the hypomethylated partition are differentially tagged.
- [0039] Embodiment 34 is the method of any one of the preceding embodiments, wherein the DNA of the hypermethylated partition and the DNA of the hypomethylated partition are differentially tagged with sequence tags comprising barcodes.
- [0040] Embodiment 35 is the method of any one of the preceding embodiments, wherein the hypermethylated and hypomethylated partitions were prepared by contacting the DNA of the sample with a methyl binding reagent immobilized on a solid support.
- [0041] Embodiment 36 is the method of the immediately preceding embodiment, wherein the methyl binding reagent comprises an MBD.
- [0042] Embodiment 37 is the method of embodiment 36, wherein the methyl binding reagent comprises an MeCP.

[0043] Embodiment 38 is the method of embodiment 36, wherein the methyl binding reagent comprises an antibody that binds a methylated nucleotide, optionally wherein the methylated nucleotide is methylated cytosine.

[0044] Embodiment 39 is the method of any one of embodiments 35-38, wherein the method comprises contacting the DNA of the sample with the methyl binding reagent immobilized on the solid support and obtaining the hypomethylated partition and hypermethylated partition based on differential binding to the methyl binding reagent.

[0045] Embodiment 40 is the method of any one of embodiments 35-39, wherein the method comprises adding differential tags to the DNA of the hypermethylated partition and the DNA of the hypomethylated partition before sequencing.

[0046] Embodiment 41 is the method of any one of the preceding embodiments, wherein determining the sequence comprises mapping the first and second sets of sequence reads to a reference sequence to produce mapped sequence reads.

[0047] Embodiment 42 is the method of any one of the preceding embodiments, wherein the DNA of the sample or of the hypermethylated and hypomethylated partitions comprises enriched or captured regions of interest.

[0048] Embodiment 43 is the method of any one of the preceding embodiments, wherein the method comprises enriching the DNA of the sample or of the hypermethylated and hypomethylated partitions for regions of interest or capturing regions of interest from the sample or the hypermethylated and hypomethylated partitions.

[0049] Embodiment 44 is the method of the immediately preceding embodiment, wherein enriching or capturing comprises contacting the DNA with a set of target-specific probes, whereby a captured set of DNA molecules is produced.

[0050] Embodiment 45 is the method of any one of embodiments 42-44, wherein the regions of interest comprise sequence-variable target regions.

[0051] Embodiment 46 is the method of the immediately preceding embodiment, wherein the set of target-specific probes comprises target-binding probes specific for a sequence-variable target set.

[0052] Embodiment 47 is the method of the immediately preceding embodiment, wherein the footprint of sequence-variable target region set is at least 25kB or at least 50kB.

- [0053] Embodiment 48 is the method of any one of embodiments 42-47, wherein the regions of interest comprise epigenetic target regions.
- [0054] Embodiment 49 is the method of the immediately preceding embodiment, wherein the set of target-specific probes comprises target-binding probes specific for an epigenetic target set.
- [0055] Embodiment 50 is the method of any one of embodiments 42-49, wherein the regions of interest comprise a sequence-variable target region set and epigenetic target region set.
- [0056] Embodiment 51 is the method of the immediately preceding embodiment, wherein there are at least 10 regions in the sequence-variable target region set and at least 100 regions in the epigenetic target region set.
- [0057] Embodiment 52 is the method of any one of embodiments 50-51, wherein the footprint of the epigenetic target region set is at least 2-fold greater than the size of the sequence-variable target region set.
- [0058] Embodiment 53 is the method of the immediately preceding embodiment, wherein the footprint of the epigenetic target region set is at least 10-fold greater than the size of the sequence-variable target region set.
- [0059] Embodiment 54 is the method of embodiment 52 or 53, wherein the set of target-specific probes is configured to capture cfDNA corresponding to the sequence-variable target set with a greater capture yield than cfDNA corresponding to the epigenetic target set.
- [0060] Embodiment 55 is the method of any one of embodiments 50-54, wherein the sequence-variable target region set has a footprint in the range of 10-30 kilobases.
- [0061] Embodiment 56 is the method of any one of embodiments 50-54, wherein the sequence-variable target region set has a footprint in the range of 30-60 kilobases.
- [0062] Embodiment 57 is the method of any one of embodiments 50-54, wherein the sequence-variable target region set has a footprint in the range of 60 kilobases to 1 megabase.
- [0063] Embodiment 58 is the method of any one of embodiments 50-54, wherein the sequence-variable target region set has a footprint in the range of 1-2 megabases.
- [0064] Embodiment 59 is the method of any one of embodiments 50-58, wherein the epigenetic target region set has a footprint in the range of 0.2-0.8 megabases.
- [0065] Embodiment 60 is the method of any one of embodiments 50-58, wherein the epigenetic target region set has a footprint in the range of 0.8-1.5 megabases.

- [0066] Embodiment 61 is the method of any one of embodiments 50-58, wherein the epigenetic target region set has a footprint in the range of 1.5-3 megabases.
- [0067] Embodiment 62 is the method of any one of embodiments 50-58, wherein the epigenetic target region set has a footprint in the range of 3-8 megabases.
- [0068] Embodiment 63 is the method of any one of embodiments 50-62, wherein the epigenetic target region set comprises a hypermethylation variable target region set.
- [0069] Embodiment 64 is the method of any one of embodiments 50-63, wherein the epigenetic target region set comprises a hypomethylation variable target region set.
- [0070] Embodiment 65 is the method of any one of embodiments 50-64, wherein the epigenetic target region set comprise a fragmentation variable target region set.
- [0071] Embodiment 66 is the method of the immediately preceding embodiment, wherein the fragmentation variable target region set comprises transcription start site regions.
- [0072] Embodiment 67 is the method of embodiment 65 or 66, wherein the fragmentation variable target region set comprises CTCF binding regions.
- [0073] Embodiment 68 is the method of any one of embodiments 50-67, wherein the captured DNA of the sequence-variable target set is sequenced to a greater depth of sequencing than the captured DNA of the epigenetic target region set.
- [0074] Embodiment 69 is the method of the immediately preceding embodiment, wherein the captured DNA of the sequence-variable target set is sequenced to at least a 2-fold, 3-fold, or 4-fold greater depth of sequencing, or is sequenced to a 4-10-fold or 4-100-fold greater depth of sequencing, than the captured cfDNA molecules of the epigenetic target region set.
- [0075] Embodiment 70 is the method of any one of embodiments 50-69, wherein the captured DNA of the sequence-variable target set are pooled with the captured DNA of the epigenetic target region set before sequencing.
- [0076] Embodiment 71 is the method of any one of embodiments 50-70, wherein the captured DNA of the sequence-variable target set and the captured DNA of the epigenetic target region set are sequenced in the same sequencing cell.
- [0077] Embodiment 72 is the method of any one of embodiments 50-71, wherein the DNA of the hypermethylated and hypomethylated partitions is amplified before capture.
- [0078] Embodiment 73 is the method of any one of the preceding embodiments, wherein the sample was obtained from a biological tissue or fluid.

[0079] Embodiment 74 is the method of any one of the preceding embodiments, wherein the sample was obtained from blood.

[0080] Embodiment 75 is the method of any one of the preceding embodiments, wherein the DNA of the sample comprises cell-free DNA.

[0081] Embodiment 76 is the method of any one of the preceding embodiments, wherein the DNA of the sample consists essentially of cell-free DNA.

[0082] Embodiment 77 is the method of any one of the preceding embodiments, wherein the sample is from a subject having or suspected of having a proliferative disorder or solid tumor.

[0083] Embodiment 78 is the method of any one of the preceding embodiments, wherein the sample is from a subject that is undergoing or has undergone treatment for a proliferative disorder or solid tumor.

[0084] Embodiment 79 is the method of any one of the preceding embodiments, further comprising determining a likelihood that the subject has a proliferative disorder or solid tumor based on the sequence determined from the sequence reads.

[0085] Embodiment 80 is the method of any one of the three immediately preceding embodiments, wherein the proliferative disorder or solid tumor is a cancer.

[0086] In some embodiments, the results of the methods disclosed herein are used as an input to generate a report. The report may be in a paper or electronic format. For example the classification of a C to T or G to A transition mutation, as obtained by the methods disclosed herein, can be displayed directly in such a report. Alternatively or additionally, diagnostic information or therapeutic recommendations based on the classification of whether a C to T or G to A transition mutation exists can be included in the report.

[0087] The various steps of the methods disclosed herein may be carried out at the same or different times, in the same or different geographical locations, e.g. countries, and/or by the same or different people.

BRIEF DESCRIPTION OF THE DRAWINGS

[0088] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate certain embodiments, and together with the written description, serve to explain certain principles of the methods, computer readable media, and systems disclosed herein. The description provided herein is better understood when read in conjunction with the

accompanying drawings which are included by way of example and not by way of limitation. It will be understood that like reference numerals identify like components throughout the drawings, unless the context indicates otherwise. It will also be understood that some or all of the figures may be schematic representations for purposes of illustration and do not necessarily depict the actual relative sizes or locations of the elements shown.

[0089] **FIG. 1** shows an overview of partitioning methodology.

[0090] **FIG. 2** is a schematic diagram of an example of a system suitable for use with some embodiments of the disclosure.

[0091] **FIG. 3** shows single nucleotide variation (SNV) per-base error rates as a function of specific nucleotide substitution.

DETAILED DESCRIPTION

[0092] Reference will now be made in detail to certain embodiments of the invention. While the invention will be described in conjunction with such embodiments, it will be understood that they are not intended to limit the invention to those embodiments. On the contrary, the invention is intended to cover all alternatives, modifications, and equivalents, which may be included within the invention as defined by the appended claims.

[0093] Before describing the present teachings in detail, it is to be understood that the disclosure is not limited to specific compositions or process steps, as such may vary. It should be noted that, as used in this specification and the appended claims, the singular form “a”, “an” and “the” include plural references unless the context clearly dictates otherwise. Thus, for example, reference to “a nucleic acid” includes a plurality of nucleic acids, reference to “a cell” includes a plurality of cells, and the like.

[0094] Numeric ranges are inclusive of the numbers defining the range. Measured and measurable values are understood to be approximate, taking into account significant digits and the error associated with the measurement. Also, the use of “comprise”, “comprises”, “comprising”, “contain”, “contains”, “containing”, “include”, “includes”, and “including” are not intended to be limiting. It is to be understood that both the foregoing general description and detailed description are exemplary and explanatory only and are not restrictive of the teachings.

[0095] Unless specifically noted in the above specification, embodiments in the specification that recite “comprising” various components are also contemplated as “consisting of” or “consisting essentially of” the recited components; embodiments in the specification that recite “consisting

of” various components are also contemplated as “comprising” or “consisting essentially of” the recited components; and embodiments in the specification that recite “consisting essentially of” various components are also contemplated as “consisting of” or “comprising” the recited components (this interchangeability does not apply to the use of these terms in the claims).

[0096] The section headings used herein are for organizational purposes and are not to be construed as limiting the disclosed subject matter in any way. In the event that any document or other material incorporated by reference contradicts any explicit content of this specification, including definitions, this specification controls.

I. DEFINITIONS

[0097] “Calling a C to T or G to A transition mutation relative to a reference sequence” refers to making a conclusion that sequence reads from a sample support the existence of a mutation at a given position in the nucleic acid being sequenced relative to the reference sequence at an acceptable level of confidence. In some embodiments, the conclusion is made computationally and/or is based on the number of reads in which the difference at the position appears, optionally in combination with other parameters such as a measure of sequence quality (e.g., Phred quality score). In some embodiments, the conclusion is made computationally and/or is based on the number of molecules in which the difference at the position appears, optionally in combination with other parameters such as a measure of mutation allele fraction and/or a background error rate, which can be estimated from, e.g., historical data or data from a control group such as a healthy cohort. In these embodiments, the molecule count is estimated using the molecular barcodes and/or genomic co-ordinates of the sequence reads. Generating an assembled sequence including the mutation and generating a report listing the mutation are non-limiting examples of calling the mutation. The assembled sequence or report may be displayed, printed, or otherwise transmitted to a user or other individual.

[0098] “Cell-free DNA,” “cfDNA molecules,” or simply “cfDNA” include DNA molecules that occur in a subject in extracellular form (e.g., in blood, serum, plasma, or other bodily fluids such as lymph, cerebrospinal fluid, urine, or sputum) and includes DNA not contained within or otherwise bound to a cell. While the DNA originally existed in a cell or cells in a large complex biological organism, e.g., a mammal, the DNA has undergone release from the cell(s) into a fluid found in the organism. Typically, cfDNA may be obtained by obtaining a sample of the fluid

without the need to perform an in vitro cell lysis step and also includes removal of cells present in the fluid (e.g., centrifugation of blood to remove cells).

[0099] The “capture yield” of a collection of probes for a given target region set refers to the amount (e.g., amount relative to another target region set or an absolute amount) of nucleic acid corresponding to the target region set that the collection captures under typical conditions. Exemplary typical capture conditions are an incubation of the sample nucleic acid and probes at 65°C for 10-18 hours in a small reaction volume (about 20 µL) containing stringent hybridization buffer. The capture yield may be expressed in absolute terms or, for a plurality of collections of probes, relative terms. When capture yields for a plurality of sets of target regions are compared, they are normalized for the footprint size of the target region set (e.g., on a per-kilobase basis). Thus, for example, if the footprint sizes of first and second target regions are 50 kb and 500 kb, respectively (giving a normalization factor of 0.1), then the DNA corresponding to the first target region set is captured with a higher yield than DNA corresponding to the second target region set when the mass per volume concentration of the captured DNA corresponding to the first target region set is more than 0.1 times the mass per volume concentration of the captured DNA corresponding to the second target region set. As a further example, using the same footprint sizes, if the captured DNA corresponding to the first target region set has a mass per volume concentration of 0.2 times the mass per volume concentration of the captured DNA corresponding to the second target region set, then the DNA corresponding to the first target region set was captured with a two-fold greater capture yield than the DNA corresponding to the second target region set.

[00100] “Capturing” or “enriching” one or more target nucleic acids refers to preferentially isolating or separating the one or more target nucleic acids from non-target nucleic acids.

[00101] A “captured set” of nucleic acids refers to nucleic acids that have undergone capture.

[00102] A “target-region set” or “set of target regions” or “target regions” refers to a plurality of genomic loci or a plurality of genomic regions targeted for capture and/or targeted by a set of probes (e.g., through sequence complementarity).

[00103] “Corresponding to a target region set” means that a nucleic acid, such as cfDNA, originated from a locus in the target region set or specifically binds one or more probes for the target-region set.

[00104] “Specifically binds” in the context of an probe or other oligonucleotide and a target sequence means that under appropriate hybridization conditions, the oligonucleotide or probe hybridizes to its target sequence, or replicates thereof, to form a stable probe:target hybrid, while at the same time formation of stable probe:non-target hybrids is minimized. Thus, a probe hybridizes to a target sequence or replicate thereof to a sufficiently greater extent than to a non-target sequence, to enable capture or detection of the target sequence. Appropriate hybridization conditions are well-known in the art, may be predicted based on sequence composition, or can be determined by using routine testing methods (see, e.g., Sambrook et al., *Molecular Cloning, A Laboratory Manual*, 2nd ed. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989) at §§ 1.90-1.91, 7.37-7.57, 9.47-9.51 and 11.47-11.57, particularly §§ 9.50-9.51, 11.12-11.13, 11.45-11.47 and 11.55-11.57, incorporated by reference herein).

[00105] “Sequence-variable target region set” refers to a set of target regions that may exhibit changes in sequence such as nucleotide substitutions, insertions, deletions, or gene fusions or transpositions in neoplastic cells (e.g., tumor cells and cancer cells).

[00106] “Epigenetic target region set” refers to a set of target regions that may manifest non-sequence modifications in neoplastic cells (e.g., tumor cells and cancer cells) and non-tumor cells (e.g., immune cells, cells from tumor microenvironment). These modifications do not change the sequence of the DNA. Examples of non-sequence modifications changes include, but not limited to, changes in methylation (increases or decreases), nucleosome distribution, CTCF binding, transcription start sites, regulatory protein binding regions and any other proteins that may bind to the DNA. For present purposes, loci susceptible to neoplasia-, tumor-, or cancer-associated focal amplifications and/or gene fusions may also be included in an epigenetic target region set because detection of a change in copy number by sequencing or a fused sequence that maps to more than one locus in a reference genome tends to be more similar to detection of exemplary epigenetic changes discussed above than detection of nucleotide substitutions, insertions, or deletions, e.g., in that the focal amplifications and/or gene fusions can be detected at a relatively shallow depth of sequencing because their detection does not depend on the accuracy of base calls at one or a few individual positions. For example, the epigenetic target region set can comprise a set of target regions for analyzing the fragment length or fragment end point location distribution.

[00107] A circulating tumor DNA or ctDNA is a component of cfDNA that originated from a tumor cell or cancer cell. In some embodiments, cfDNA comprises DNA that originated from normal

cells and DNA that originated from tumor cells (i.e., ctDNA). Tumor cells are neoplastic cells that originated from a tumor, regardless of whether they remain in the tumor or become separated from the tumor (as in the cases, e.g., of metastatic cancer cells and circulating tumor cells).

[00108] The term “hypermethylation” refers to an increased level or degree of methylation of nucleic acid molecule(s) relative to the other nucleic acid molecules within a population (e.g., sample) of nucleic acid molecules. In some embodiments, hypermethylated DNA can include DNA molecules comprising at least 1 methylated residue, at least 2 methylated residues, at least 3 methylated residues, at least 5 methylated residues, at least 10 methylated residues, at least 20 methylated residues, at least 25 methylated residues, or at least 30 methylated residues.

[00109] The term “hypomethylation” refers to a decreased level or degree of methylation of nucleic acid molecule(s) relative to the other nucleic acid molecules within a population (e.g., sample) of nucleic acid molecules. In some embodiments, hypomethylated DNA includes unmethylated DNA molecules. In some embodiments, hypomethylated DNA can include DNA molecules comprising 0 methylated residues, at most 1 methylated residue, at most 2 methylated residues, at most 3 methylated residues, at most 4 methylated residues, or at most 5 methylated residues.

[00110] The term “methylated nucleotide” refers to a nucleotide to which a methyl group is attached, other than the methyl attached to the pyrimidine ring of thymine. An example of a methylated nucleotide is a nucleotide comprising 5-methylcytosine or 7-methylguanine.

[00111] As used herein, a “sequence of a molecule” and grammatical variants thereof refer to a sequence determined from a plurality of reads comprising reads from molecules derived from the same original sample molecule. Reads may be determined to be derived from the same original sample molecule based on, e.g., one or more of the sequences of a tag or barcode; the genomic positions corresponding to the first and last nucleotides of the sample sequence; and/or the sequence(s) of a plurality of bases immediately following a 5' tag sequence and/or immediately preceding a 3' tag sequence. In some embodiments, each base in a sequence of a molecule is determined based on agreement of a minimum number of reads at that position, e.g., at least 2, 3, 4, 5, 6, 7, 8, 9, or 10 reads; the number of reads required may depend on whether a call is supported by reads of both strands of DNA or only one strand, for example, the number of reads required may increase by 1, 2, 3, 4, or 5 reads if there are reads for only one strand of the sequence relative to if there are reads for both strands of the sequence of the molecule.

[00112] The terms “or a combination thereof” and “or combinations thereof” as used herein refers to any and all permutations and combinations of the listed terms preceding the term. For example, “A, B, C, or combinations thereof” is intended to include at least one of: A, B, C, AB, AC, BC, or ABC, and if order is important in a particular context, also BA, CA, CB, ACB, CBA, BCA, BAC, or CAB. Continuing with this example, expressly included are combinations that contain repeats of one or more item or term, such as BB, AAA, AAB, BBC, AAABCCCC, CBBAAA, CABABB, and so forth. The skilled artisan will understand that typically there is no limit on the number of items or terms in any combination, unless otherwise apparent from the context.

[00113] “Or” is used in the inclusive sense, i.e., equivalent to “and/or,” unless the context requires otherwise.

II. EXEMPLARY METHODS

[00114] Provided herein are methods of analyzing a sample of DNA. In some embodiments, the methods comprise obtaining first and second sets of sequence reads from the hypermethylated and hypomethylated partitions, respectively. In some embodiments, the methods comprise obtaining first and second sets of sequences of molecules from the hypermethylated and hypomethylated partitions, respectively. The sequences of molecules can be obtained, for example, by partitioning the DNA sample into a plurality of partitions, wherein the plurality of partitions comprises a hypermethylated partition and a hypomethylated partition; tagging the DNA in the hypermethylated and hypomethylated partitions to generate tagged nucleic acids, wherein the tagged nucleic acids comprise molecular barcodes; obtaining sequence reads of molecules from the hypermethylated partition and sequence reads of molecules from the hypomethylated partition, wherein the sequence reads comprise molecular barcode sequence and sample sequence; and grouping sequence reads into families based on at least one of (a) the molecular barcode sequences and (b) genomic positions corresponding to the first and last nucleotides of the sample sequence, wherein the families comprise sequence reads derived from a single DNA molecule in the sample. In some embodiments, the methods comprise determining a sequence from the first and second sets of sequence reads or sequences of molecules. The term “sequence” is used in a collective sense and does not necessarily imply one continuous sequence. That is, it may refer to a whole genome sequence (e.g., comprising multiple chromosome sequences), a set of genomic loci or genes, any other set of sequences, the identity of a base at an individual position, or combinations thereof. In some embodiments, the methods comprise determining a first set of sequences of

molecules from the hypermethylated partition and a second set of sequences of molecules from the hypomethylated partition. In some embodiments, the methods comprise obtaining a set of sequences of molecules from the hypomethylated partition. The sequences of molecules can be obtained, for example, by partitioning the DNA sample into a plurality of partitions, wherein the plurality of partitions comprises a hypermethylated partition and a hypomethylated partition; tagging the DNA in the hypomethylated partition to generate tagged nucleic acids, wherein the tagged nucleic acids comprise molecular barcodes; obtaining sequence reads of molecules from the hypomethylated partition, wherein the sequence reads comprise molecular barcode sequence and sample sequence; and grouping sequence reads into families based on at least one of (a) the molecular barcode sequences and (b) genomic positions corresponding to the first and last nucleotides of the sample sequence, wherein the families comprise sequence reads derived from a single DNA molecule in the sample. In some embodiments, the methods comprise determining a sequence from the set of sequence reads or sequences of molecules.

[00115] The methods may comprise calling a C to T or G to A transition mutation relative to a reference sequence. The reference sequence may be a standard genome sequence for the organism from which the sample was obtained (e.g., a mammal, such as a human). Alternatively, the reference sequence may be another sequence from the same subject from whom the sample was obtained; in such cases the reference sequence may be from, e.g., healthy tissue or an earlier time point.

[00116] In some embodiments, calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set requires observation of the transition mutation in a greater number of reads than calling a C to T or G to A transition mutation relative to the reference sequence based on reads of the second set. In some embodiments, calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set requires observation of the transition mutation in a greater number of sequences of molecules than calling a C to T or G to A transition mutation relative to the reference sequence based on sequences of molecules of the second set. As described elsewhere herein, it has been determined that hypermethylated partitions (e.g., of cfDNA) frequently contain more damaged (e.g., deaminated) DNA than hypomethylated partitions, which results in sequence reads with apparent C to T or G to A transition mutations that do not correspond to the actual sequence in the cell from which the DNA originated. Deaminated bases that do not correspond to an actual in vivo mutation

relative to the reference sequence may be referred to as artifactual deaminations. Without wishing to be bound by any particular theory, damaged (e.g., deaminated) DNA may be preferentially partitioned into the hypermethylated partition, e.g., because it is possible that hypermethylated DNA may be more susceptible to damage such as deamination or more exposed to damaging agents such as deaminating agents. As such, there is likely to be an increased risk of artifactual deaminations and therefore calling false positive C to T or G to A transition mutations when sequencing DNA from a hypermethylated partition. Requiring observation of transition mutations in a greater number of reads or molecules to call a C to T or G to A transition mutation relative to a reference sequence based on reads or molecules of the first set can compensate for the increased frequency of artifactual deaminations and reduce or eliminate the risk of elevated false positive calls of transition mutations in the sequence determined from the hypermethylated partition.

[00117] For example, calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set may require observation of the transition mutation in at least three reads. In such embodiments, calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set may require observation of the transition mutation in one or two reads, e.g., in two reads. In some embodiments, calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set may require observation of the transition mutation in at least a portion of reads that include the base. The portion may be, e.g., three reads per 10,000 reads that include the base, or at least 0.1%, 0.5%, 1%, 2%, 5%, 10%, 20%, 30% or 40% of the reads that include the base. In some embodiments, the portion of reads may be less than 0.1% of the reads that include the base. In some embodiments, the portion of reads may be at least 40% of the reads that include the base. Optionally, calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the second set may require observation of the transition mutation in one or two reads per 10,000 reads that include the base, e.g., two reads per 10,000 reads that include the base, or at least 0.1%, 0.5%, 1%, 2%, 5%, 10%, 20%, or 30% of the reads that include the base, wherein the number of required observations is lower than the number required for calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set.

[00118] In another example, calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set may require observation of the transition mutation in at least three molecules. In such embodiments, calling a C to T or G to A transition

mutation relative to a reference sequence based on sequences of molecules of the first set may require observation of the transition mutation in one or two molecules, e.g., in two molecules. In some embodiments, calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set may require observation of the transition mutation in at least three sequences of molecules per 10,000 molecules that include the base, or at least 0.1%, 0.5%, 1%, 2%, 5%, 10%, 20%, 30% or 40% of the sequences of molecules that include the base. Optionally, calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set may require observation of the transition mutation in one or two sequences of molecules per 10,000 molecules that include the base, e.g., in two sequences of molecules per 10,000 molecules that include the base, or at least three per 10,000, 0.1%, 0.5%, 1%, 2%, 5%, 10%, 20%, or 30% of the sequences of molecules that include the base, wherein the number of required observations is lower than the number required for calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set.

[00119] In another example, calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set may require observation of the transition mutation in at least four reads. In such embodiments, calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set may require observation of the transition mutation in one, two, or three reads, e.g., in two or three reads.

[00120] In another example, calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set may require observation of the transition mutation in at least four molecules. In such embodiments, calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set may require observation of the transition mutation in one, two, or three molecules, e.g., in two or three molecules.

[00121] In another example, calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set may require observation of the transition mutation in at least five reads. In such embodiments, calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set may require observation of the transition mutation in one, two, three, or four reads, e.g., in two, three, or four reads, or more particularly two or three reads.

[00122] In another example, calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set may require observation of the transition mutation in at least five molecules. In such embodiments, calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set may require observation of the transition mutation in one, two, three, or four molecules, e.g., in two, three, or four molecules, or more particularly two or three molecules. More generally, the number of observations required for calling a C to T or G to A transition mutation may be as shown in the following Table 1.

[00123] Table 1. Exemplary parameters for calling C to T or G to A transition mutation

Number of sequences of molecules to call C to T or G to A transition mutation in first set of reads	Number of sequences of molecules to call C to T or G to A transition mutation in first set of reads
At least 2	1
At least 3	1 or 2
At least 4	1, 2, or 3
At least 5	1, 2, 3, or 4
At least 6	1, 2, 3, 4 or 5
At least 7	1, 2, 3, 4, 5 or 6
At least 8	1, 2, 3, 4, 5, 6 or 7

[00124] Appropriate values may be selected based on one or more, or all, of the quality of the sample, the depth of the sequence data, and the relative importance of specificity (avoiding false positives) and sensitivity (avoiding false negatives). In some embodiments, calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set requires observation of the transition mutation in at least two more reads than does calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the second set. In some embodiments, calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set requires observation of the transition mutation in at least two more molecules than does calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the second set.

[00125] In some embodiments, C to T or G to A transition mutations are not called relative to a reference sequence based on reads of the first set. In some embodiments, C to T or G to A transition mutations are not called relative to a reference sequence based on sequences of molecules of the first set. In some embodiments, C to T or G to A transition mutations are called relative to a reference sequence based on reads of the second set without the use of reads of the first set. In some embodiments, C to T or G to A transition mutations are called relative to a reference sequence based on sequences of molecules of the second set without the use of sequences of molecules of the first set. For example, to the extent C to T or G to A transition mutations are called, they may be called only based on evidence from the hypomethylated partition, or from the hypomethylated partition and one or more intermediate partitions as discussed elsewhere herein. Such embodiments eliminate the risk of calling false positive C to T or G to A transition mutations due to damaged (e.g., deaminated) DNA in the hypermethylated partition. In some embodiments, a C to T or G to A transition mutation is called relative to a reference sequence only if at least one sequence of a molecule of the second set (e.g., at least two sequences of molecules of the second set) comprises the C to T or G to A transition mutation. In some embodiments, a C to T or G to A transition mutation is called relative to a reference sequence only if at least one read of the second set (e.g., at least two reads of the second set) comprises the C to T or G to A transition mutation.

[00126] In some embodiments, a third set of sequence reads from an intermediate partition is obtained, from which a third set of sequences of molecules from the intermediate partition may be determined. In some of these embodiments, C to T or G to A transition mutations may be called based on sequences of molecules of the third set less stringently than C to T and G to A transition mutations are called based on sequences of molecules of the first set. In some of these embodiments, C to T and G to A transition mutations may be called based on sequences of molecules of the third set in the same way as C to T and G to A transition mutations are called based on sequences of molecules of the second set. In some embodiments, C to T and G to A transition mutations may be called based on sequences of molecules of the third set more stringently than C to T and G to A transition mutations are called based on sequences of molecules of the second set.

[00127] In some embodiments, a third set of sequence reads from an intermediate partition is obtained, from which a third set of sequences reads from the intermediate partition may be determined in addition to the first and second sets. In some embodiments, C to T or G to A

transition mutations may be called based on reads of the third set less stringently than C to T and G to A transition mutations are called based on reads of the first set. In some embodiments, C to T and G to A transition mutations are called based on reads of the third set in the same way as C to T and G to A transition mutations are called based on reads of the second set. In some embodiments, C to T and G to A transition mutations are called based on reads of the third set in more stringently than C to T and G to A transition mutations are called based on reads of the second set.

[00128] In some embodiments, a threshold may be used for calling a C to T or G to A transition based on sequences of molecules. For example, in some embodiments, a first threshold is used for calling a C to T or G to A transition based on sequences of molecules of the first set, and a second threshold is used for calling a C to T or G to A transition based on sequences of molecules of the second set. In some embodiments, the first threshold provides a first level of specificity for calling a C to T or G to A transition and the second threshold provides a second level of specificity for calling a C to T or G to A transition. In some embodiments, the first level of specificity is approximately equal to the second level of specificity. In other embodiments, the first level of specificity is within 10%, 5%, 4%, 3%, 2%, 1%, 0.5%, 0.2%, or 0.1% of the second level of specificity. In some embodiments, the first and second thresholds are specific for C to T and/or G to A transitions.

[00129] In some embodiments, the first and second thresholds may be determined from a plurality of control samples. In some embodiments, the first and second thresholds are determined from at least one control sample. In some of these embodiments, the control samples may be from individuals who are not suspected of having cancer.

[00130] In some embodiments, background sequencing error rates may be incorporated into the methods of the disclosure. For example, a first group of position-specific background error rates may be used for a plurality of positions for sequences of the first set of sequence of the first set. Some examples further include a second group of position-specific background error rates used for a plurality of positions for sequence of the second set. In these examples, the second group comprises position-specific background error rates higher than the corresponding position-specific background error rates of the first group. In some of these embodiments, calling a C to T or G to A transition mutation based on sequences of molecules of the first set requires observation of the

C to T or G to A transition mutation at a frequency exceeding the corresponding rate from the first group of position-specific background error rates.

[00131] In some embodiments, calling a C to T or G to A transition mutation based on sequences of molecules of the first set requires observation of the C to T or G to A transition mutation at a frequency exceeding the corresponding rate from the first group of position-specific background error rates by a factor of at least 2, 3, 4, or 5.

[00132] In some embodiments, calling a C to T or G to A transition mutation based on sequences of molecules of the first set requires observation of the C to T or G to A transition mutation at a frequency exceeding the corresponding rate from the first group of position-specific background error rates by an amount consistent with a confidence level of at least 95%, 98%, 99%, 99.5%, or 99.9%. The confidence level may be determined based on appropriate statistics, such as using statistical measures that may include standard deviation, standard error of the mean, confidence interval, t-score and Z-score. In some embodiments, the first and second groups of position-specific background error rates were determined from a plurality of control samples. In some embodiments, the control samples may be from individuals not suspected of having cancer. In some embodiments, the first and second groups of position-specific background error rates were determined using historical data, e.g., the frequency of apparent mutations that did not meet a predetermined confidence threshold in a set of previously obtained sequence data. In some embodiments, the first and second groups of position-specific background error rates were determined using reads and/or sequences of molecules from the hypermethylated and hypomethylated partitions, respectively, e.g., at runtime.

1. Partitioning; Analysis of epigenetic characteristics

[00133] In certain embodiments described herein, the methods comprise partitioning the sample of DNA, e.g., to provide the hypermethylated and hypomethylated partitions, and optionally one or more additional (e.g., intermediately methylated) partitions and/or sub-partitions of the hypermethylated and hypomethylated partitions. In general, DNA in a sample, such as a captured set of cfDNA as described elsewhere herein, can be physically partitioned based on one or more characteristics of the nucleic acids (such as methylation) prior to analysis, e.g., sequencing, or tagging and sequencing. This approach can be used to determine, for example, whether hypermethylation variable epigenetic target regions show hypermethylation characteristic of

tumor cells or hypomethylation variable epigenetic target regions show hypomethylation characteristic of tumor cells. Additionally, by partitioning a heterogeneous nucleic acid population, one may increase rare signals, e.g., by enriching rare nucleic acid molecules that are more prevalent in one fraction (or partition) of the population. For example, a genetic variation present in hyper-methylated DNA but less (or not) in hypomethylated DNA (e.g., a genetic variation other than a C to T or G to A transition mutation) can be more easily detected by partitioning a sample into hyper-methylated and hypo-methylated nucleic acid molecules. By analyzing multiple fractions of a sample, a multi-dimensional analysis of a single locus of a genome or species of nucleic acid can be performed and hence, greater sensitivity can be achieved.

[00134] In some instances, a heterogeneous nucleic acid sample is partitioned into two or more partitions (e.g., at least 3, 4, 5, 6 or 7 partitions). In some embodiments, each partition is differentially tagged. Tagged partitions can then be pooled together for collective sample prep and/or sequencing. The partitioning-tagging-pooling steps can occur more than once, with each round of partitioning occurring based on a different characteristics (examples provided herein) and tagged using differential tags that are distinguished from other partitions and partitioning means.

[00135] Examples of characteristics that can be used for partitioning include sequence length, methylation level, nucleosome binding, sequence mismatch, immunoprecipitation, and/or proteins that bind to DNA. Resulting partitions can include one or more of the following nucleic acid forms: single-stranded DNA (ssDNA), double-stranded DNA (dsDNA), shorter DNA fragments, and longer DNA fragments. In some embodiments, a heterogeneous population of nucleic acids is partitioned into nucleic acids with one or more epigenetic modifications and without the one or more epigenetic modifications. Examples of epigenetic modifications include presence or absence of methylation; level of methylation; type of methylation (e.g., 5-methylcytosine versus other types of methylation, such as adenine methylation and/or cytosine hydroxymethylation); and association and level of association with one or more proteins, such as histones. Alternatively, or additionally, a heterogeneous population of nucleic acids can be partitioned into nucleic acid molecules associated with nucleosomes and nucleic acid molecules devoid of nucleosomes. Alternatively, or additionally, a heterogeneous population of nucleic acids may be partitioned into single-stranded DNA (ssDNA) and double-stranded DNA (dsDNA). Alternatively, or additionally, a heterogeneous population of nucleic acids may be partitioned based on nucleic acid length (e.g., molecules of up to 160 bp and molecules having a length of greater than 160 bp).

[00136] In some instances, each partition (representative of a different nucleic acid form) is differentially labelled, and the partitions are pooled together prior to sequencing. In other instances, the different forms are separately sequenced.

[00137] FIG. 1 illustrates an exemplary scheme comprising partitioning. A population of different nucleic acids (101) is partitioned (102) into two or more different partitions (103 a, b). Each partition (103 a, b) is representative of a different nucleic acid form. Each partition is distinctly tagged (104). The tagged nucleic acids are pooled together (107) prior to sequencing (108). Reads are analyzed, in silico. Tags are used to sort reads from different partitions. Analysis to detect genetic variants can be performed on a partition-by-partition level, as well as whole nucleic acid population level. Apparent C to T or G to A transition mutations may be analyzed separately for a hypermethylated partition using more stringent parameters, as described in detail elsewhere herein, or reads or sequences of molecules from the hypermethylated partition may simply not be used to call C to T or G to A transition mutations. An exemplary analysis can include in silico analysis to determine genetic variants, such as CNV, SNV, indel, fusion in nucleic acids in each partition. In some instances, in silico analysis can include determining chromatin structure. For example, coverage of sequence reads can be used to determine nucleosome positioning in chromatin. Higher coverage can correlate with higher nucleosome occupancy in genomic region while lower coverage can correlate with lower nucleosome occupancy or nucleosome depleted region (NDR).

[00138] Samples can include nucleic acids varying in modifications including post-replication modifications to nucleotides and binding, usually noncovalently, to one or more proteins.

[00139] Any type of sample described elsewhere herein may be used. In an embodiment, the population of nucleic acids is one obtained from a tissue, serum, plasma or blood sample from a subject suspected of having neoplasia, a tumor, or cancer or previously diagnosed with neoplasia, a tumor, or cancer. The population of nucleic acids includes nucleic acids having varying levels of methylation. Methylation can occur from any one or more post-replication or transcriptional modifications. Post-replication modifications include modifications of the nucleotide cytosine, particularly at the 5-position of the nucleobase, e.g., 5-methylcytosine, 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine.

[00140] In some embodiments, the nucleic acids in the original population can be single-stranded and/or double-stranded. Partitioning based on single v. double stranded-ness of the nucleic acids

can be accomplished by, e.g. using labelled capture probes to partition ssDNA and using double stranded adapters to partition dsDNA.

[00141] Partitioning may be performed using any appropriate reagent, e.g., any of the reagents described elsewhere herein, to selectively bind or separate nucleic acids based on a difference in characteristics. The reagents can be antibodies with the desired specificity, natural binding partners or variants thereof (Bock et al., Nat Biotech 28: 1106-1114 (2010); Song et al., Nat Biotech 29: 68-72 (2011)), or artificial peptides selected e.g., by phage display to have specificity to a given target.

[00142] Examples of reagents contemplated herein include methyl binding domain (MBDs) and methyl binding proteins (MBPs) as described herein.

[00143] Likewise, partitioning of different forms of nucleic acids can be performed using histone binding proteins which can separate nucleic acids bound to histones from free or unbound nucleic acids. Examples of histone binding proteins that can be used in the methods disclosed herein include RBBP4 (RbAp48) and SANT domain peptides.

[00144] Although for some reagents and modifications, binding to the reagent may occur in an essentially all or none manner depending on whether a nucleic acid bears a modification, the separation may be one of degree. In such instances, nucleic acids overrepresented in a modification bind to the reagent at a greater extent than nucleic acids underrepresented in the modification. Alternatively, nucleic acids having modifications may bind in an all or nothing manner. But then, various levels of modifications may be sequentially eluted from the binding agent.

[00145] For example, in some embodiments, partitioning can be binary or based on degree/level of modifications. For example, all methylated fragments can be partitioned from unmethylated fragments using methyl-binding domain proteins (e.g., MethylMiner Methylated DNA Enrichment Kit (Thermo Fisher Scientific)). Subsequently, additional partitioning may involve eluting fragments having different levels of methylation by adjusting the salt concentration in a solution with the methyl-binding domain and bound fragments. As salt concentration increases, fragments having greater methylation levels are eluted.

[00146] In some instances, the final partitions are representatives of nucleic acids having different extents of modifications (overrepresentative or underrepresentative of modifications). Overrepresentation and underrepresentation can be defined by the number of modifications borne by a nucleic acid relative to the median number of modifications per strand in a population. For

example, if the median number of 5-methylcytosine residues in nucleic acid in a sample is 2, a nucleic acid including more than two 5-methylcytosine residues is overrepresented in this modification and a nucleic acid with 1 or zero 5-methylcytosine residues is underrepresented. The effect of the affinity separation is to enrich for nucleic acids overrepresented in a modification in a bound phase and for nucleic acids underrepresented in a modification in an unbound phase (i.e. in solution). The nucleic acids in the bound phase can be eluted before subsequent processing.

[00147] When using MethylMiner Methylated DNA Enrichment Kit (Thermo Fisher Scientific) various levels of methylation can be partitioned using sequential elutions. For example, a hypomethylated partition (e.g., no methylation) can be separated from a methylated partition by contacting the nucleic acid population with the MBD from the kit, which is attached to magnetic beads. The beads are used to separate out the methylated nucleic acids from the non-methylated nucleic acids. Subsequently, one or more elution steps are performed sequentially to elute nucleic acids having different levels of methylation. For example, a first set of methylated nucleic acids can be eluted at a salt concentration of 160 mM or higher, e.g., at least 200 mM, 300 mM, 400 mM, 500 mM, 600 mM, 700 mM, 800 mM, 900 mM, 1000 mM, or 2000 mM. After such methylated nucleic acids are eluted, magnetic separation is once again used to separate higher level of methylated nucleic acids from those with lower level of methylation. The elution and magnetic separation steps can repeat themselves to create various partitions such as a hypomethylated partition (e.g., representative of no methylation), a methylated partition (representative of low level of methylation), and a hyper methylated partition (representative of high level of methylation).

[00148] In some methods, nucleic acids bound to an agent used for affinity separation are subjected to a wash step. The wash step washes off nucleic acids weakly bound to the affinity agent. Such nucleic acids can be enriched in nucleic acids having the modification to an extent close to the mean or median (i.e., intermediate between nucleic acids remaining bound to the solid phase and nucleic acids not binding to the solid phase on initial contacting of the sample with the agent).

[00149] The affinity separation results in at least two, and sometimes three or more partitions of nucleic acids with different extents of a modification. While the partitions are still separate, the nucleic acids of at least one partition, and usually two or three (or more) partitions are linked to nucleic acid tags, usually provided as components of adapters, with the nucleic acids in different partitions receiving different tags that distinguish members of one partition from another. The tags linked to nucleic acid molecules of the same partition can be the same or different from one

another. But if different from one another, the tags may have part of their code in common so as to identify the molecules to which they are attached as being of a particular partition.

[00150] For further details regarding portioning nucleic acid samples based on characteristics such as methylation, see WO2018/119452, which is incorporated herein by reference.

[00151] In some embodiments, the nucleic acid molecules can be fractionated into different partitions based on the nucleic acid molecules that are bound to a specific protein or a fragment thereof and those that are not bound to that specific protein or fragment thereof.

[00152] Nucleic acid molecules can be fractionated based on DNA-protein binding. Protein-DNA complexes can be fractionated based on a specific property of a protein. Examples of such properties include various epitopes, modifications (e.g., histone methylation or acetylation) or enzymatic activity. Examples of proteins which may bind to DNA and serve as a basis for fractionation may include, but are not limited to, protein A and protein G. Any suitable method can be used to fractionate the nucleic acid molecules based on protein bound regions. Examples of methods used to fractionate nucleic acid molecules based on protein bound regions include, but are not limited to, SDS-PAGE, chromatin-immuno-precipitation (ChIP), heparin chromatography, and asymmetrical field flow fractionation (AF4).

[00153] In some embodiments, partitioning of the nucleic acids is performed by contacting the nucleic acids with a methylation binding domain (“MBD”) of a methylation binding protein (“MBP”). MBD binds to 5-methylcytosine (5mC). MBD is coupled to paramagnetic beads, such as Dynabeads® M-280 Streptavidin via a biotin linker. Partitioning into fractions with different extents of methylation can be performed by eluting fractions by increasing the NaCl concentration.

[00154] Examples of MBPs contemplated herein include, but are not limited to:

- (a) MeCP2 is a protein preferentially binding to 5-methyl-cytosine over unmodified cytosine.
- (b) RPL26, PRP8 and the DNA mismatch repair protein MHS6 preferentially bind to 5- hydroxymethyl-cytosine over unmodified cytosine.
- (c) FOXK1, FOXK2, FOXP1, FOXP4 and FOXI3 preferably bind to 5-formyl-cytosine over unmodified cytosine (Iurlaro et al., Genome Biol. 14: R119 (2013)).
- (d) Antibodies specific to one or more methylated nucleotide bases.

[00155] In general, elution is a function of number of methylated sites per molecule, with molecules having more methylation eluting under increased salt concentrations. To elute the DNA

into distinct populations based on the extent of methylation, one can use a series of elution buffers of increasing NaCl concentration. Salt concentration can range from about 100 mM to about 2500 mM NaCl. In one embodiment, the process results in three (3) partitions. Molecules are contacted with a solution at a first salt concentration and comprising a molecule comprising a methyl binding domain, which molecule can be attached to a capture moiety, such as streptavidin. At the first salt concentration a population of molecules will bind to the MBD and a population will remain unbound. The unbound population can be separated as a “hypomethylated” population. For example, a first partition representative of the hypomethylated form of DNA is that which remains unbound at a low salt concentration, e.g., 100 mM or 160 mM. A second partition representative of intermediate methylated DNA is eluted using an intermediate salt concentration, e.g., between 100 mM and 2000 mM concentration. This is also separated from the sample. A third partition representative of hypermethylated form of DNA is eluted using a high salt concentration, e.g., at least about 2000 mM.

a. Tagging of partitions

[00156] In some embodiments, two or more partitions of the sample of DNA, e.g., each partition, have been or are differentially tagged. Tags can be molecules, such as nucleic acids, containing information that indicates a feature of the molecule with which the tag is associated. For example, molecules can bear a sample tag (which distinguishes molecules in one sample from those in a different sample), a partition tag (which distinguishes molecules in one partition from those in a different partition) or a molecular tag (which distinguishes different molecules from one another (in both unique and non-unique tagging scenarios). In certain embodiments, a tag can comprise one or a combination of barcodes. As used herein, the term “barcode” refers to a nucleic acid molecule having a particular nucleotide sequence, or to the nucleotide sequence, itself, depending on context. A barcode can have, for example, between 10 and 100 nucleotides. A collection of barcodes can have degenerate sequences or can have sequences having a certain Hamming distance, as desired for the specific purpose. So, for example, a sample index, partition index or molecular index can be comprised of one barcode or a combination of two barcodes, each attached to different ends of a molecule.

[00157] Tags can be used to label the individual polynucleotide population partitions so as to correlate the tag (or tags) with a specific partition. Alternatively, tags can be used in embodiments of the invention that do not employ a partitioning step. In some embodiments, a single tag can be

used to label a specific partition. In some embodiments, multiple different tags can be used to label a specific partition. In embodiments employing multiple different tags to label a specific partition, the set of tags used to label one partition can be readily differentiated for the set of tags used to label other partitions. In some embodiments, the tags may have additional functions, for example the tags can be used to index sample sources or used as unique molecular identifiers (which can be used to improve the quality of sequencing data by differentiating sequencing errors from mutations, for example as in Kinde et al., *Proc Nat'l Acad Sci USA* 108: 9530-9535 (2011), Kou et al., *PLoS ONE*, 11: e0146638 (2016)) or used as non-unique molecule identifiers, for example as described in US Pat. No. 9,598,731. Similarly, in some embodiments, the tags may have additional functions, for example the tags can be used to index sample sources or used as non-unique molecular identifiers (which can be used to improve the quality of sequencing data by differentiating sequencing errors from mutations).

[00158] In one embodiment, partition tagging comprises tagging molecules in each partition with a partition tag. After re-combining partitions and sequencing molecules, the partition tags identify the source partition. In another embodiment, different partitions are tagged with different sets of molecular tags, e.g., comprised of a pair of barcodes. In this way, each molecular barcode indicates the source partition as well as being useful to distinguish molecules within a partition. For example, a first set of 35 barcodes can be used to tag molecules in a first partition, while a second set of 35 barcodes can be used tag molecules in a second partition.

[00159] In some embodiments, after partitioning and tagging with partition tags, the molecules may be pooled for sequencing in a single run. In some embodiments, a sample tag is added to the molecules, e.g., in a step subsequent to addition of partition tags and pooling. Sample tags can facilitate pooling material generated from multiple samples for sequencing in a single sequencing run.

[00160] Alternatively, in some embodiments, partition tags may be correlated to the sample as well as the partition. As a simple example, a first tag can indicate a first partition of a first sample; a second tag can indicate a second partition of the first sample; a third tag can indicate a first partition of a second sample; and a fourth tag can indicate a second partition of the second sample.

[00161] While tags may be attached to molecules already partitioned based on one or more characteristics, the final tagged molecules in the library may no longer possess that characteristic. For example, while single stranded DNA molecules may be partitioned and tagged, the final tagged

molecules in the library are likely to be double stranded. Similarly, while DNA may be subject to partition based on different levels of methylation, in the final library, tagged molecules derived from these molecules are likely to be unmethylated. Accordingly, the tag attached to molecule in the library typically indicates the characteristic of the “parent molecule” from which the ultimate tagged molecule is derived, not necessarily to characteristic of the tagged molecule, itself.

[00162] As an example, barcodes 1, 2, 3, 4, etc. are used to tag and label molecules in the first partition; barcodes A, B, C, D, etc. are used to tag and label molecules in the second partition; and barcodes a, b, c, d, etc. are used to tag and label molecules in the third partition. Differentially tagged partitions can be pooled prior to sequencing. Differentially tagged partitions can be separately sequenced or sequenced together concurrently, e.g., in the same flow cell of an Illumina sequencer.

[00163] After sequencing, analysis of reads to detect genetic variants can be performed on a partition-by-partition level, as well as a whole nucleic acid population level. Tags are used to sort reads from different partitions. Analysis can include in silico analysis to determine genetic and epigenetic variation (one or more of methylation, chromatin structure, etc.) using sequence information, genomic coordinates length, coverage and/or copy number. In some embodiments, higher coverage can correlate with higher nucleosome occupancy in genomic region while lower coverage can correlate with lower nucleosome occupancy or a nucleosome depleted region (NDR).

b. Determination of 5-methylcytosine pattern of nucleic acids; Bisulfite sequencing

[00164] Bisulfite-based sequencing and variants thereof provides a means of determining the methylation pattern of a nucleic acid that can provide single-base resolution information regarding methylation status. In some embodiments, determining the methylation pattern comprises distinguishing 5-methylcytosine (5mC) from non-methylated cytosine. In some embodiments, determining methylation pattern comprises distinguishing N-methyladenine from non-methylated adenine. In some embodiments, determining the methylation pattern comprises distinguishing 5-hydroxymethyl cytosine (5hmC), 5-formyl cytosine (5fC), and 5- carboxylcytosine (5caC) from non-methylated cytosine. Examples of bisulfite sequencing include but are not limited to oxidative bisulfite sequencing (OX-BS-seq), Tet-assisted bisulfite sequencing (TAB-seq), and reduced bisulfite sequencing (redBS-seq).

[00165] Oxidative bisulfite sequencing (OX-BS-seq) is used to distinguish between 5mC and 5hmC, by first converting the 5hmC to 5fC, and then proceeding with bisulfite sequencing. Tet-assisted bisulfite sequencing (TAB-seq) can also be used to distinguish 5mc and 5hmC. In TAB-seq, 5hmC is protected by glucosylation. A Tet enzyme is then used to convert 5mC to 5caC before proceeding with bisulfite sequencing. Reduced bisulfite sequencing is used to distinguish 5fC from modified cytosines.

[00166] Generally, in bisulfite sequencing, a nucleic acid sample is divided into two aliquots and one aliquot is treated with bisulfite. In some embodiments, a hypermethylated partition is divided into two such aliquots. The bisulfite converts native cytosine and certain modified cytosine nucleotides (e.g. 5-formyl cytosine or 5-carboxylcytosine) to uracil whereas other modified cytosines (e.g., 5-methylcytosine, 5-hydroxymethylcystosine) are not converted. Comparison of nucleic acid sequences of molecules from the two aliquots indicates which cytosines were and were not converted to uracils. Consequently, cytosines which were and were not modified can be determined. The initial splitting of the sample into two aliquots is disadvantageous for samples containing only small amounts of nucleic acids, and/or composed of heterogeneous cell/tissue origins such as bodily fluids containing cell-free DNA.

[00167] Accordingly, in some embodiments, bisulfite sequencing is performed without initially splitting a sample into two aliquots, e.g., as follows. In some embodiments, nucleic acids in a population are linked to a capture moiety such as any of those described herein, i.e., a label that can be captured or immobilized. Following linking of capture moieties to sample nucleic acids, the sample nucleic acids serve as templates for amplification. Following amplification, the original templates remain linked to the capture moieties but amplicons are not linked to capture moieties.

[00168] The capture moiety can be linked to sample nucleic acids as a component of an adapter, which may also provide amplification and/or sequencing primer binding sites. In some methods, sample nucleic acids are linked to adapters at both ends, with both adapters bearing a capture moiety. Preferably any cytosine residues in the adapters are modified, such as by 5-methylation, to protect against the action of bisulfite. In some instances, the capture moieties are linked to the original templates by a cleavable linkage (e.g., photocleavable desthiobiotin-TEG or uracil residues cleavable with USERTM enzyme, Chem. Commun. (Camb). 51: 3266-3269 (2015)), in which case the capture moieties can, if desired, be removed.

[00169] The amplicons are denatured and contacted with an affinity reagent for the capture tag. Original templates bind to the affinity reagent whereas nucleic acid molecules resulting from amplification do not. Thus, the original templates can be separated from nucleic acid molecules resulting from amplification.

[00170] Following separation of original templates from nucleic acid molecules resulting from amplification, the original templates can be subjected to bisulfite treatment. Alternatively, the amplification products can be subjected to bisulfite treatment and the original template population not. Following such treatment, the respective populations can be amplified (which in the case of the original template population converts uracils to thymines). The populations can also be subjected to biotin probe hybridization for capture. The respective populations are then analyzed and sequences compared to determine which cytosines were 5- methylated (or 5-hydroxymethylated) in the original sample. Detection of a T nucleotide in the template population (corresponding to an unmethylated cytosine converted to uracil) and a C nucleotide at the corresponding position of the amplified population indicates an unmodified C. The presence of C's at corresponding positions of the original template and amplified populations indicates a modified C in the original sample.

[00171] In some embodiments, a method uses sequential DNA-seq and bisulfite-seq (BIS- seq) NGS library preparation of molecular tagged DNA libraries (see WO 2018/119452, e.g., at FIG. 4). This process is performed by labeling of adapters (e.g., biotin), DNA-seq amplification of whole library, parent molecule recovery (e.g. streptavidin bead pull down), bisulfite conversion and BIS-seq. In some embodiments, the method identifies 5-methylcytosine with single-base resolution, through sequential NGS-preparative amplification of parent library molecules with and without bisulfite treatment. This can be achieved by modifying the 5-methylated NGS-adapters (directional adapters; Y-shaped/forked with 5-methylcytosine replacing) used in BIS-seq with a label (e.g., biotin) on one of the two adapter strands. Sample DNA molecules are adapter ligated, and amplified (e.g., by PCR). As only the parent molecules will have a labeled adapter end, they can be selectively recovered from their amplified progeny by label-specific capture methods (e.g., streptavidin-magnetic beads). As the parent molecules retain 5-methylation marks, bisulfite conversion on the captured library will yield single-base resolution 5-methylation status upon BIS-seq, retaining molecular information to corresponding DNA-seq. In some embodiments, the bisulfite treated library can be combined with a non-treated library prior to capture/NGS by

addition of a sample tag DNA sequence in standard multiplexed NGS workflow. As with BIS-seq workflows, bioinformatics analysis can be carried out for genomic alignment and 5-methylated base identification. In sum, this method provides the ability to selectively recover the parent, ligated molecules, carrying 5-methylcytosine marks, after library amplification, thereby allowing for parallel processing for bisulfite converted DNA. This overcomes the destructive nature of bisulfite treatment on the quality/sensitivity of the DNA-seq information extracted from a workflow. With this method, the recovered ligated, parent DNA molecules (via labeled adapters) allow amplification of the complete DNA library and parallel application of treatments that elicit epigenetic DNA modifications. The present disclosure discusses the use of BIS-seq methods to identify cytosine-5-methylation (5-methylcytosine), but the use of BIS-seq methods is not required in many embodiments. Variants of BIS-seq have been developed to identify hydroxymethylated cytosines (5hmC; OX-BS-seq, TAB-seq), formylcytosine (5fC; redBS-seq) and carboxyl cytosines. These methodologies can be implemented with the sequential/parallel library preparation described herein.

c. Alternative Methods of Modified Nucleic Acid Analysis

[00172] In some such methods, a population of nucleic acids bearing the modification to different extents (e.g., 0, 1, 2, 3, 4, 5 or more methyl groups per nucleic acid molecule) is contacted with adapters before fractionation of the population depending on the extent of the modification. Adapters attach to either one end or both ends of nucleic acid molecules in the population. Preferably, the adapters include different tags of sufficient numbers that the number of combinations of tags results in a low probability e.g., 95, 99 or 99.9% of two nucleic acids with the same start and stop points receiving the same combination of tags. Following attachment of adapters, the nucleic acids are amplified from primers binding to the primer binding sites within the adapters. Adapters, whether bearing the same or different tags (e.g., tags of the same or different sequence), can include the same or different primer binding sites, but preferably adapters include the same primer binding site. Following amplification, the nucleic acids are contacted with an agent that preferably binds to nucleic acids bearing the modification (such as the previously described such agents). The nucleic acids are separated into at least two partitions (e.g., hypermethylated and hypomethylated partitions) differing in the extent to which the nucleic acids bear the modification from binding to the agents. For example, if the agent has affinity for nucleic acids bearing the modification, nucleic acids overrepresented in the modification (compared with

median representation in the population) preferentially bind to the agent, whereas nucleic acids underrepresented for the modification do not bind or are more easily eluted from the agent. Following separation, the different partitions can then be subject to further processing steps, which typically include further amplification, and sequence analysis as described elsewhere herein, in parallel but separately. Sequence data from the different partitions can then be compared.

[00173] Such a separation scheme can be performed using the following exemplary procedure. Nucleic acids are linked at both ends to Y-shaped adapters including primer binding sites and tags. The molecules are amplified. The amplified molecules are then fractionated by contact with an antibody preferentially binding to 5-methylcytosine to produce two partitions. One partition includes original molecules lacking methylation and amplification copies having lost methylation. The other partition includes original DNA molecules with methylation. The two partitions are then processed and sequenced separately with further amplification of the methylated partition. The sequence data of the two partitions can then be compared. In this example, tags are not used to distinguish between methylated and unmethylated DNA but rather to distinguish between different molecules within these partitions so that one can determine whether reads with the same start and stop points are based on the same or different molecules.

[00174] Methods described herein may further comprise analyzing a population of nucleic acid (e.g., a hypermethylated partition) in which at least some of the nucleic acids include one or more modified cytosine residues, such as 5-methylcytosine and any of the other modifications described previously. In these methods, the population of nucleic acids is contacted with adapters including one or more cytosine residues modified at the 5C position, such as 5-methylcytosine. Preferably all cytosine residues in such adapters are also modified, or all such cytosines in a primer binding region of the adapters are modified. Adapters attach to both ends of nucleic acid molecules in the population. Preferably, the adapters include different tags of sufficient numbers that the number of combinations of tags results in a low probability e.g., 95, 99 or 99.9% of two nucleic acids with the same start and stop points receiving the same combination of tags. The primer binding sites in such adapters can be the same or different, but are preferably the same. After attachment of adapters, the nucleic acids are amplified from primers binding to the primer binding sites of the adapters. The amplified nucleic acids are split into first and second aliquots. The first aliquot is assayed for sequence data with or without further processing. The sequence data on molecules in the first aliquot is thus determined irrespective of the initial methylation state of the nucleic acid

molecules. The nucleic acid molecules in the second aliquot are treated with bisulfite. This treatment converts unmodified cytosines to uracils. The bisulfite treated nucleic acids are then subjected to amplification primed by primers to the original primer binding sites of the adapters linked to nucleic acid. Only the nucleic acid molecules originally linked to adapters (as distinct from amplification products thereof) are now amplifiable because these nucleic acids retain cytosines in the primer binding sites of the adapters, whereas amplification products have lost the methylation of these cytosine residues, which have undergone conversion to uracils in the bisulfite treatment. Thus, only original molecules in the populations, at least some of which are methylated, undergo amplification. After amplification, these nucleic acids are subject to sequence analysis. Comparison of sequences determined from the first and second aliquots can indicate among other things, which cytosines in the nucleic acid population were subject to methylation.

[00175] Such an analysis can be performed using the following exemplary procedure. Methylated DNA is linked to Y-shaped adapters at both ends including primer binding sites and tags. The cytosines in the adapters are 5-methylated. The methylation of the primers serves to protect the primer binding sites in a subsequent bisulfite step. After attachment of adapters, the DNA molecules are amplified. The amplification product is split into two aliquots for sequencing with and without bisulfite treatment. The aliquot not subjected to bisulfite sequencing can be subjected to sequence analysis with or without further processing. The other aliquot is treated with bisulfite, which converts unmethylated cytosines to uracils. Only primer binding sites protected by methylation of cytosines can support amplification when contacted with primers specific for original primer binding sites. Thus, only original molecules and not copies from the first amplification are subjected to further amplification. The further amplified molecules are then subjected to sequence analysis. Sequences can then be compared from the two aliquots. As in the separation scheme discussed above, nucleic acid tags in adapters are not used to distinguish between methylated and unmethylated DNA but to distinguish nucleic acid molecules within the same partition.

2. Target regions; differential capture and sequencing depth

[00176] In some embodiments, the methods comprise capturing cfDNA obtained from a test subject for a plurality of sets of target regions. The target regions comprise epigenetic target regions, which may show differences in methylation levels and/or fragmentation patterns depending on whether

they originated from a tumor or from healthy cells. The target regions also comprise sequence-variable target regions, which may show differences in sequence depending on whether they originated from a tumor or from healthy cells. The capturing step produces a captured set of cfDNA molecules, and the cfDNA molecules corresponding to the sequence-variable target region set are captured at a greater capture yield in the captured set of cfDNA molecules than cfDNA molecules corresponding to the epigenetic target region set.

[00177] In some embodiments, the methods comprise contacting cfDNA obtained from a test subject with a set of target-specific probes, wherein the set of target-specific probes is configured to capture cfDNA corresponding to the sequence-variable target region set at a greater capture yield than cfDNA corresponding to the epigenetic target region set.

[00178] It can be beneficial to capture cfDNA corresponding to the sequence-variable target region set at a greater capture yield than cfDNA corresponding to the epigenetic target region set because a greater depth of sequencing may be necessary to analyze the sequence-variable target regions with sufficient confidence or accuracy than may be necessary to analyze the epigenetic target regions. The greater depth of sequencing can result in more reads per DNA molecule and can be facilitated by capturing more unique molecules per region. The volume of data needed to determine fragmentation patterns (e.g., to test for perturbation of transcription start sites or CTCF binding sites) or fragment abundance (e.g., in hypermethylated and hypomethylated partitions) is generally less than the volume of data needed to determine the presence or absence of cancer-related sequence mutations. Capturing the target region sets at different yields can facilitate sequencing the target regions to different depths of sequencing in the same sequencing run (e.g., using a pooled mixture and/or in the same sequencing cell).

[00179] In various embodiments, the methods further comprise sequencing the captured cfDNA, e.g., to different degrees of sequencing depth for the epigenetic and sequence-variable target region sets, consistent with the discussion above.

a. Captured set(s); differential capture and sequencing depth

[00180] In some embodiments, a captured set of DNA (e.g., cfDNA) is provided. With respect to the disclosed methods, the captured set of DNA may be provided, e.g., following capturing, and/or partitioning steps as described herein. The captured set may comprise DNA corresponding to a sequence-variable target region set and an epigenetic target region set. In some embodiments the quantity of captured sequence-variable target region DNA is greater than the quantity of the

captured epigenetic target region DNA, when normalized for the difference in the size of the targeted regions (footprint size).

[00181] Alternatively, first and second captured sets may be provided, comprising, respectively, DNA corresponding to a sequence-variable target region set and DNA corresponding to an epigenetic target region set. The first and second captured sets may be combined to provide a combined captured set.

[00182] In a captured set comprising DNA corresponding to the sequence-variable target region set and the epigenetic target region set, including a combined captured set as discussed above, the DNA corresponding to the sequence-variable target region set may be present at a greater concentration than the DNA corresponding to the epigenetic target region set, e.g., a 1.1 to 1.2-fold greater concentration, a 1.2- to 1.4-fold greater concentration, a 1.4- to 1.6-fold greater concentration, a 1.6- to 1.8-fold greater concentration, a 1.8- to 2.0-fold greater concentration, a 2.0- to 2.2-fold greater concentration, a 2.2- to 2.4-fold greater concentration a 2.4- to 2.6-fold greater concentration, a 2.6- to 2.8-fold greater concentration, a 2.8- to 3.0-fold greater concentration, a 3.0- to 3.5-fold greater concentration, a 3.5- to 4.0, a 4.0- to 4.5-fold greater concentration, a 4.5- to 5.0-fold greater concentration, a 5.0- to 5.5-fold greater concentration, a 5.5- to 6.0-fold greater concentration, a 6.0- to 6.5-fold greater concentration, a 6.5- to 7.0-fold greater, a 7.0- to 7.5-fold greater concentration, a 7.5- to 8.0-fold greater concentration, an 8.0- to 8.5-fold greater concentration, an 8.5- to 9.0-fold greater concentration, a 9.0- to 9.5-fold greater concentration, 9.5- to 10.0-fold greater concentration, a 10- to 11-fold greater concentration, an 11- to 12-fold greater concentration a 12- to 13-fold greater concentration, a 13- to 14-fold greater concentration, a 14- to 15-fold greater concentration, a 15- to 16-fold greater concentration, a 16- to 17-fold greater concentration, a 17- to 18-fold greater concentration, an 18- to 19-fold greater concentration, or a 19- to 20-fold greater concentration. The degree of difference in concentrations accounts for normalization for the footprint sizes of the target regions, as discussed in the definition section.

i. Epigenetic target region set

[00183] The epigenetic target region set may comprise one or more types of target regions likely to differentiate DNA from neoplastic (e.g., tumor or cancer) cells and from healthy cells, e.g., non-neoplastic circulating cells. Exemplary types of such regions are discussed in detail herein. The

epigenetic target region set may also comprise one or more control regions, e.g., as described herein.

[00184] In some embodiments, the epigenetic target region set has a footprint of at least 100 kb, e.g., at least 200 kb, at least 300 kb, or at least 400 kb. In some embodiments, the epigenetic target region set has a footprint in the range of 100-1000 kb, e.g., 100-200 kb, 200-300 kb, 300-400 kb, 400-500 kb, 500-600 kb, 600-700 kb, 700-800 kb, 800-900 kb, and 900-1,000 kb. In some embodiments, the epigenetic target region set has a footprint of at least 1000 kb, at least 2000 kb, at least 3000 kb, at least 4000 kb, at least 5000 kb, at least 6000 kb, at least 7000 kb, at least 8000 kb, at least 9000 kb or at least 1 Mb. In some embodiments, the epigenetic target region set has a footprint in the range of 1 Mb-20 Mb, e.g., 1-1.2 Mb, 1.2-1.4 Mb, 1.4-1.6 Mb, 1.6-1.8 Mb, 1.8-2 Mb, 2-2.25 Mb, 2.25-2.5 Mb, 2.5-2.75 Mb, 2.75-3 Mb, 3-3.25 Mb, 3.25-3.5 Mb, 3.5-3.75 Mb, 3.75-4 Mb, 4-4.25 Mb, 4.25-4.5 Mb, 4.5-4.75 Mb, 4.75-5 Mb, 5-5.5 Mb, 5.5-6 Mb, 6-6.5 Mb, 6.5-7 Mb, 7-7.5 Mb, 7.5-8 Mb, 8-8.5 Mb, 8.5-9 Mb, 9-9.5 Mb, 9.5-10 Mb, 10-12 Mb, 12-14 Mb, 14-16 Mb, 16-18 Mb, and 18-20 Mb. In some embodiments, the epigenetic target region set has a footprint in the range of 0.2-0.8 megabases, 0.8-1.5 megabases, 1.5-3 megabases, or 3-8 megabases.

(a) Hypermethylation variable target regions

[00185] In some embodiments, the epigenetic target region set comprises one or more hypermethylation variable target regions. In general, hypermethylation variable target regions refer to regions where an increase in the level of observed methylation indicates an increased likelihood that a sample (e.g., of cfDNA) contains DNA produced by neoplastic cells, such as tumor or cancer cells. For example, hypermethylation of promoters of tumor suppressor genes has been observed repeatedly. See, e.g., Kang et al., *Genome Biol.* 18:53 (2017) and references cited therein.

[00186] An extensive discussion of methylation variable target regions in colorectal cancer is provided in Lam et al., *Biochim Biophys Acta.* 1866:106-20 (2016). These include VIM, SEPT9, ITGA4, OSM4, GATA4 and NDRG4. An exemplary set of hypermethylation variable target regions comprising the genes or portions thereof based on the colorectal cancer (CRC) studies is provided in Table 2A. Many of these genes likely have relevance to cancers beyond colorectal cancer; for example, TP53 is widely recognized as a critically important tumor suppressor and hypermethylation-based inactivation of this gene may be a common oncogenic mechanism.

Table 2A. Exemplary hypermethylation target regions (genes or portions thereof) based on CRC studies.

Gene Name	Additional Gene Name	Chromosome
VIM		chr10
SEPT9		chr17
CYCD2	CCND2	chr12
TFPI2		chr7
GATA4		chr8
RARB2	RARB	chr3
p16INK4a	CDKN2A	chr9
MGMT	MGMT	chr10
APC		chr5
NDRG4		chr16
HLTF		chr3
HPP1	TMEFF2	chr2
hMLH1	MLH1	chr3
RASSF1A	RASSF1	chr3
CDH13		chr16
IGFBP3		chr7
ITGA4		chr2

[00187] In some embodiments, the hypermethylation variable target regions comprise a plurality of genes or portions thereof listed in Table 2A, e.g., at least 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 100% of the genes or portions thereof listed in Table 2A. For example, for each locus included as a target region, there may be one or more probes with a hybridization site that binds between the transcription start site and the stop codon (the last stop codon for genes that are alternatively spliced) of the gene. In some embodiments, the one or more probes bind within 300 bp upstream and/or downstream of the genes or portions thereof listed in Table 2A, e.g., within 200 or 100 bp.

[00188] Methylation variable target regions in various types of lung cancer are discussed in detail, e.g., in Ooki et al., *Clin. Cancer Res.* 23:7141-52 (2017); Belinsky, *Annu. Rev. Physiol.* 77:453-74 (2015); Hulbert et al., *Clin. Cancer Res.* 23:1998-2005 (2017); Shi et al., *BMC Genomics* 18:901 (2017); Schneider et al., *BMC Cancer* 11:102 (2011); Lissa et al., *Transl Lung Cancer Res* 5(5):492-504 (2016); Skvortsova et al., *Br. J. Cancer* 94(10):1492–1495 (2006); Kim et al., *Cancer Res.* 61:3419–3424 (2001); Furonaka et al., *Pathology International* 55:303-309 (2005); Gomes et al., *Rev. Port. Pneumol.* 20:20-30 (2014); Kim et al., *Oncogene* 20:1765-70 (2001); Hopkins-Donaldson et al., *Cell Death Differ.* 10:356-64 (2003); Kikuchi et al., *Clin. Cancer Res.* 11:2954-61 (2005); Heller et al., *Oncogene* 25:959–968 (2006); Licchesi et al., *Carcinogenesis* 29:895–904 (2008); Guo et al., *Clin. Cancer Res.* 10:7917-24 (2004); Palmisano et al., *Cancer Res.* 63:4620–4625 (2003); and Toyooka et al., *Cancer Res.* 61:4556–4560, (2001).

[00189] An exemplary set of hypermethylation variable target regions comprising genes or portions thereof based on the lung cancer studies is provided in Table 2B. Many of these genes likely have relevance to cancers beyond lung cancer; for example, Casp8 (Caspase 8) is a key enzyme in programmed cell death and hypermethylation-based inactivation of this gene may be a common oncogenic mechanism not limited to lung cancer. Additionally, a number of genes appear in both Tables 2A and 2B, indicating generality.

Table 2B. Exemplary hypermethylation target regions (genes or portions thereof) based on lung cancer studies

Gene Name	Chromosome
MARCH11	chr5
TAC1	chr7
TCF21	chr6
SHOX2	chr3
p16	chr3
Casp8	chr2
CDH13	chr16
MGMT	chr10
MLH1	chr3
MSH2	chr2

TSLC1	chr11
APC	chr5
DKK1	chr10
DKK3	chr11
LKB1	chr11
WIF1	chr12
RUNX3	chr1
GATA4	chr8
GATA5	chr20
PAX5	chr9
E-Cadherin	chr16
H-Cadherin	chr16

[00190] Any of the foregoing embodiments concerning target regions identified in Table 2B may be combined with any of the embodiments described above concerning target regions identified in Table 2A. In some embodiments, the hypermethylation variable target regions comprise a plurality of genes or portions thereof listed in Table 2A or Table 2B, e.g., at least 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 100% of the genes or portions thereof listed in Table 2A or Table 2B.

[00191] Additional hypermethylation target regions may be obtained, e.g., from the Cancer Genome Atlas. Kang et al., *Genome Biology* 18:53 (2017), describe construction of a probabilistic method called Cancer Locator using hypermethylation target regions from breast, colon, kidney, liver, and lung. In some embodiments, the hypermethylation target regions can be specific to one or more types of cancer. Accordingly, in some embodiments, the hypermethylation target regions include one, two, three, four, or five subsets of hypermethylation target regions that collectively show hypermethylation in one, two, three, four, or five of breast, colon, kidney, liver, and lung cancers.

(b) Hypomethylation variable target regions

[00192] Global hypomethylation is a commonly observed phenomenon in various cancers. See, e.g., Hon et al., *Genome Res.* 22:246-258 (2012) (breast cancer); Ehrlich, *Epigenomics* 1:239-259 (2009) (review article noting observations of hypomethylation in colon, ovarian, prostate,

leukemia, hepatocellular, and cervical cancers). For example, regions such as repeated elements, e.g., LINE1 elements, Alu elements, centromeric tandem repeats, pericentromeric tandem repeats, and satellite DNA, and intergenic regions that are ordinarily methylated in healthy cells may show reduced methylation in tumor cells. Accordingly, in some embodiments, the epigenetic target region set includes hypomethylation variable target regions, where a decrease in the level of observed methylation indicates an increased likelihood that a sample (e.g., of cfDNA) contains DNA produced by neoplastic cells, such as tumor or cancer cells.

[00193] In some embodiments, hypomethylation variable target regions include repeated elements and/or intergenic regions. In some embodiments, repeated elements include one, two, three, four, or five of LINE1 elements, Alu elements, centromeric tandem repeats, pericentromeric tandem repeats, and/or satellite DNA.

[00194] Exemplary specific genomic regions that show cancer-associated hypomethylation include nucleotides 8403565-8953708 and 151104701-151106035 of human chromosome 1, e.g., according to the hg19 human genome construct. In some embodiments, the hypomethylation variable target regions overlap or comprise one or both of these regions.

(c) CTCF binding regions

[00195] CTCF is a DNA-binding protein that contributes to chromatin organization and often colocalizes with cohesin. Perturbation of CTCF binding sites has been reported in a variety of different cancers. *See, e.g.,* Katainen et al., *Nature Genetics*, doi:10.1038/ng.3335, published online 8 June 2015; Guo et al., *Nat. Commun.* 9:1520 (2018). CTCF binding results in recognizable patterns in cfDNA that can be detected by sequencing, e.g., through fragment length analysis. For example, details regarding sequencing-based fragment length analysis are provided in Snyder et al., *Cell* 164:57-68 (2016); WO 2018/009723; and US20170211143A1, each of which are incorporated herein by reference.

[00196] Thus, perturbations of CTCF binding result in variation in the fragmentation patterns of cfDNA. As such, CTCF binding sites represent a type of fragmentation variable target regions.

[00197] There are many known CTCF binding sites. *See, e.g.,* the CTCFBSDB (CTCF Binding Site Database), available on the Internet at insulatordb.uthsc.edu/; Cuddapah et al., *Genome Res.* 19:24-32 (2009); Martin et al., *Nat. Struct. Mol. Biol.* 18:708-14 (2011); Rhee et al., *Cell.* 147:1408-19 (2011), each of which are incorporated by reference. Exemplary CTCF binding sites

are at nucleotides 56014955-56016161 on chromosome 8 and nucleotides 95359169-95360473 on chromosome 13, e.g., according to the hg19 or hg38 human genome construct.

[00198] Accordingly, in some embodiments, the epigenetic target region set includes CTCF binding regions. In some embodiments, the CTCF binding regions comprise at least 10, 20, 50, 100, 200, or 500 CTCF binding regions, or 10-20, 20-50, 50-100, 100-200, 200-500, or 500-1000 CTCF binding regions, e.g., such as CTCF binding regions described above or in one or more of CTCFBSDB or the Cuddapah et al., Martin et al., or Rhee et al. articles cited above.

[00199] In some embodiments, at least some of the CTCF sites can be methylated or unmethylated, wherein the methylation state is correlated with the whether or not the cell is a cancer cell. In some embodiments, the epigenetic target region set comprises at least 100 bp, at least 200 bp, at least 300 bp, at least 400 bp, at least 500 bp, at least 750 bp, at least 1000 bp upstream and/or downstream regions of the CTCF binding sites.

(d) Transcription start sites

[00200] Transcription start sites may also show perturbations in neoplastic cells. For example, nucleosome organization at various transcription start sites in healthy cells of the hematopoietic lineage—which contributes substantially to cfDNA in healthy individuals—may differ from nucleosome organization at those transcription start sites in neoplastic cells. This results in different cfDNA patterns that can be detected by sequencing, for example, as discussed generally in Snyder et al., *Cell* 164:57-68 (2016); WO 2018/009723; and US20170211143A1.

[00201] Thus, perturbations of transcription start sites also result in variation in the fragmentation patterns of cfDNA. As such, transcription start sites also represent a type of fragmentation variable target regions.

[00202] Human transcriptional start sites are available from DBTSS (DataBase of Human Transcription Start Sites), available on the Internet at dbtss.hgc.jp and described in Yamashita et al., *Nucleic Acids Res.* 34(Database issue): D86–D89 (2006), which is incorporated herein by reference.

[00203] Accordingly, in some embodiments, the epigenetic target region set includes transcriptional start sites. In some embodiments, the transcriptional start sites comprise at least 10, 20, 50, 100, 200, or 500 transcriptional start sites, or 10-20, 20-50, 50-100, 100-200, 200-500, or 500-1000 transcriptional start sites, e.g., such as transcriptional start sites listed in DBTSS. In some embodiments, at least some of the transcription start sites can be methylated or unmethylated,

wherein the methylation state is correlated with the whether or not the cell is a cancer cell. In some embodiments, the epigenetic target region set comprises at least 100 bp, at least 200 bp, at least 300 bp, at least 400 bp, at least 500 bp, at least 750 bp, at least 1000 bp upstream and/or downstream regions of the transcription start sites.

(e) Focal amplifications

[00204] Although focal amplifications are somatic mutations, they can be detected by sequencing based on read frequency in a manner analogous to approaches for detecting certain epigenetic changes such as changes in methylation. As such, regions that may show focal amplifications in cancer can be included in the epigenetic target region set and may comprise one or more of AR, BRAF, CCND1, CCND2, CCNE1, CDK4, CDK6, EGFR, ERBB2, FGFR1, FGFR2, KIT, KRAS, MET, MYC, PDGFRA, PIK3CA, and RAF1. For example, in some embodiments, the epigenetic target region set comprises at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, or 18 of the foregoing targets.

(f) Methylation control regions

[00205] It can be useful to include control regions to facilitate data validation. In some embodiments, the epigenetic target region set includes control regions that are expected to be methylated or unmethylated in essentially all samples, regardless of whether the DNA is derived from a cancer cell or a normal cell. In some embodiments, the epigenetic target region set includes control hypomethylated regions that are expected to be hypomethylated in essentially all samples. In some embodiments, the epigenetic target region set includes control hypermethylated regions that are expected to be hypermethylated in essentially all samples.

ii. Sequence-variable target region set

[00206] In some embodiments, the sequence-variable target region set comprises a plurality of regions known to undergo somatic mutations in cancer.

[00207] In some embodiments, the sequence-variable target region set targets a plurality of different genes or genomic regions ("panel") selected such that a determined proportion of subjects having a cancer exhibits a genetic variant or tumor marker in one or more different genes or genomic regions in the panel. The panel may be selected to limit a region for sequencing to a fixed number of base pairs. The panel may be selected to sequence a desired amount of DNA, e.g., by adjusting the affinity and/or amount of the probes as described elsewhere herein. The panel may be further

selected to achieve a desired sequence read depth. The panel may be selected to achieve a desired sequence read depth or sequence read coverage for an amount of sequenced base pairs. The panel may be selected to achieve a theoretical sensitivity, a theoretical specificity, and/or a theoretical accuracy for detecting one or more genetic variants in a sample.

[00208] Probes for detecting the panel of regions can include those for detecting genomic regions of interest (hotspot regions) as well as nucleosome-aware probes (e.g., KRAS codons 12 and 13) and may be designed to optimize capture based on analysis of cfDNA coverage and fragment size variation impacted by nucleosome binding patterns and GC sequence composition. Regions used herein can also include non-hotspot regions optimized based on nucleosome positions and GC models.

[00209] Examples of listings of genomic locations of interest may be found in Table 3 and Table 4. In some embodiments, a sequence-variable target region set used in the methods of the present disclosure comprises at least a portion of 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 70 of the genes of Table 3. In some embodiments, a sequence-variable target region set used in the methods of the present disclosure comprises 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 70 of the SNVs of Table 3. In some embodiments, a sequence-variable target region set used in the methods of the present disclosure comprises at least 1, at least 2, at least 3, at least 4, at least 5, or 6 of the fusions of Table 3. In some embodiments, a sequence-variable target region set used in the methods of the present disclosure comprise at least a portion of at least 1, at least 2, or 3 of the indels of Table 3. In some embodiments, a sequence-variable target region set used in the methods of the present disclosure comprises at least a portion of 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, at least 70, or 73 of the genes of Table 4. In some embodiments, a sequence-variable target region set used in the methods of the present disclosure comprises 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, at least 70, or 73 of the SNVs of Table 4. In some embodiments, a sequence-variable target region set used in the methods of the present disclosure comprises at least 1, at least 2, at least 3, at least 4, at least 5, or 6 of the fusions of Table 4. In some embodiments, a sequence-variable target region set used in the methods of the present disclosure comprises at least

a portion of at least 1, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, or 18 of the indels of Table 4. Each of these genomic locations of interest may be identified as a backbone region or hot-spot region for a given panel. An example of a listing of hot-spot genomic locations of interest may be found in Table 5. In some embodiments, a sequence-variable target region set used in the methods of the present disclosure comprises at least a portion of at least 1, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, or at least 20 of the genes of Table 5. Each hot-spot genomic region is listed with several characteristics, including the associated gene, chromosome on which it resides, the start and stop position of the genome representing the gene's locus, the length of the gene's locus in base pairs, the exons covered by the gene, and the critical feature (e.g., type of mutation) that a given genomic region of interest may seek to capture.

Table 3

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

Table 4

Point Mutations (SNVs) and Indels						Fusions
AKT1	ALK	APC	AR	ARAF	ARID1A	ALK
ATM	BRAF	BRCA1	BRCA2	CCND1	CCND2	FGFR2
CCNE1	CDH1	CDK4	CDK6	CDKN2A	DDR2	FGFR3
CTNNB1	EGFR	ERBB2	ESR1	EZH2	FBXW7	NTRK1
FGFR1	FGFR2	FGFR3	GATA3	GNA11	GNAQ	RET
GNAS	HNF1A	HRAS	IDH1	IDH2	JAK2	ROS1
JAK3	KIT	KRAS	MAP2K1	MAP2K2	MET	
MLH1	MPL	MYC	NF1	NFE2L2	NOTCH1	
NPM1	NRAS	NTRK1	PDGFRA	PIK3CA	PTEN	
PTPN11	RAF1	RB1	RET	RHEB	RHOA	

RIT1	ROS1	SMAD4	SMO	MAPK1	STK11	
TERT	TP53	TSC1	VHL	MAPK3	MTOR	
NTRK3						

Table 5

Gene	Chromosome	Start Position	Stop Position	Length (bp)	Exons Covered	Feature
ALK	chr2	29446405	29446655	250	intron 19	Fusion
ALK	chr2	29446062	29446197	135	intron 20	Fusion
ALK	chr2	29446198	29446404	206	20	Fusion
ALK	chr2	29447353	29447473	120	intron 19	Fusion
ALK	chr2	29447614	29448316	702	intron 19	Fusion
ALK	chr2	29448317	29448441	124	19	Fusion
ALK	chr2	29449366	29449777	411	intron 18	Fusion
ALK	chr2	29449778	29449950	172	18	Fusion
BRAF	chr7	140453064	140453203	139	15	BRAF V600
CTNNB1	chr3	41266007	41266254	247	3	S37
EGFR	chr7	55240528	55240827	299	18 and 19	G719 and deletions
EGFR	chr7	55241603	55241746	143	20	Insertions/T790M
EGFR	chr7	55242404	55242523	119	21	L858R
ERBB2	chr17	37880952	37881174	222	20	Insertions
ESR1	chr6	152419857	152420111	254	10	V534, P535, L536, Y537, D538
FGFR2	chr10	123279482	123279693	211	6	S252
GATA3	chr10	8111426	8111571	145	5	SS / Indels
GATA3	chr10	8115692	8116002	310	6	SS / Indels
GNAS	chr20	57484395	57484488	93	8	R844
IDH1	chr2	209113083	209113394	311	4	R132
IDH2	chr15	90631809	90631989	180	4	R140, R172
KIT	chr4	55524171	55524258	87	1	
KIT	chr4	55561667	55561957	290	2	
KIT	chr4	55564439	55564741	302	3	
KIT	chr4	55565785	55565942	157	4	
KIT	chr4	55569879	55570068	189	5	
KIT	chr4	55573253	55573463	210	6	

KIT	chr4	55575579	55575719	140	7	
KIT	chr4	55589739	55589874	135	8	
KIT	chr4	55592012	55592226	214	9	
KIT	chr4	55593373	55593718	345	10 and 11	557, 559, 560, 576
KIT	chr4	55593978	55594297	319	12 and 13	V654
KIT	chr4	55595490	55595661	171	14	T670, S709
KIT	chr4	55597483	55597595	112	15	D716
KIT	chr4	55598026	55598174	148	16	L783
KIT	chr4	55599225	55599368	143	17	C809, R815, D816, L818, D820, S821F, N822, Y823
KIT	chr4	55602653	55602785	132	18	A829P
KIT	chr4	55602876	55602996	120	19	
KIT	chr4	55603330	55603456	126	20	
KIT	chr4	55604584	55604733	149	21	
KRAS	chr12	25378537	25378717	180	4	A146
KRAS	chr12	25380157	25380356	199	3	Q61
KRAS	chr12	25398197	25398328	131	2	G12/G13
MET	chr7	116411535	116412255	720	13, 14, intron 13, intron 14	MET exon 14 SS
NRAS	chr1	115256410	115256609	199	3	Q61
NRAS	chr1	115258660	115258791	131	2	G12/G13
PIK3CA	chr3	178935987	178936132	145	10	E545K
PIK3CA	chr3	178951871	178952162	291	21	H1047R
PTEN	chr10	89692759	89693018	259	5	R130
SMAD4	chr18	48604616	48604849	233	12	D537
TERT	chr5	1294841	1295512	671	promoter	chr5:1295228
TP53	chr17	7573916	7574043	127	11	Q331, R337, R342
TP53	chr17	7577008	7577165	157	8	R273
TP53	chr17	7577488	7577618	130	7	R248
TP53	chr17	7578127	7578299	172	6	R213/Y220
TP53	chr17	7578360	7578564	204	5	R175 / Deletions
TP53	chr17	7579301	7579600	299	4	

				12574 (total target region)		
				16330 (total probe coverage)		

[00210] Additionally, or alternatively, suitable target region sets are available from the literature. For example, Gale et al., PLoS One 13: e0194630 (2018), which is incorporated herein by reference, describes a panel of 35 cancer-related gene targets that can be used as part or all of a sequence-variable target region set. These 35 targets are AKT1, ALK, BRAF, CCND1, CDK2A, CTNNB1, EGFR, ERBB2, ESR1, FGFR1, FGFR2, FGFR3, FOXL2, GATA3, GNA11, GNAQ, GNAS, HRAS, IDH1, IDH2, KIT, KRAS, MED12, MET, MYC, NFE2L2, NRAS, PDGFRA, PIK3CA, PPP2R1A, PTEN, RET, STK11, TP53, and U2AF1.

[00211] In some embodiments, the sequence-variable target region set comprises target regions from at least 10, 20, 30, or 35 cancer-related genes, such as the cancer-related genes listed above. In some embodiments, the sequence-variable target region set has a footprint of at least 10 kb, at least 20 kb, at least 30 kb, at least 40 kb, at least 50 kb, at least 75 kb, at least 100 kb, at least 200 kb, at least 300 kb, or at least 400 kb. In some embodiments, the sequence-variable target region set has a footprint in the range of 100-1000 kb, e.g., 100-200 kb, 200-300 kb, 300-400 kb, 400-500 kb, 500-600 kb, 600-700 kb, 700-800 kb, 800-900 kb, and 900-1000 kb. In some embodiments, the sequence-variable target region set has a footprint of at least 1000 kb, at least 2000 kb, at least 3000 kb, at least 4000 kb, at least 5000 kb, at least 6000 kb, at least 7000 kb, at least 8000 kb, at least 9000 kb or at least 1 Mb. In some embodiments, the sequence-variable target region set has a footprint in the range of 1 Mb-10 Mb, e.g., 1-1.2 Mb, 1.2-1.4 Mb, 1.4-1.6 Mb, 1.6-1.8 Mb, 1.8-2 Mb, 2-2.25 Mb, 2.25-2.5 Mb, 2.5-2.75 Mb, 2.75-3 Mb, 3-3.25 Mb, 3.25-3.5 Mb, 3.5-3.75 Mb, 3.75-4 Mb, 4-4.25 Mb, 4.25-4.5 Mb, 4.5-4.75 Mb, 4.75-5 Mb, 5-5.5 Mb, 5.5-6 Mb, 6-6.5 Mb, 6.5-7 Mb, 7-7.5 Mb, 7.5-8 Mb, 8-8.5 Mb, 8.5-9 Mb, 9-9.5 Mb, and 9.5-10 Mb. In some embodiments, the sequence-variable target region set has a footprint in the range of 10-30 kilobases, 30-60 kilobases, 60 kilobases to 1 megabase, or 1-2 megabases.

3. Subjects; sample type/source

[00212] In some embodiments, the DNA (e.g., cfDNA or DNA obtained from tissue sample) is obtained from a subject having a cancer. In some embodiments, the DNA (e.g., cfDNA or DNA obtained from tissue sample) is obtained from a subject suspected of having a cancer. In some embodiments, the DNA (e.g., cfDNA or DNA obtained from tissue sample) is obtained from a subject having a tumor. In some embodiments, the DNA (e.g., cfDNA or DNA obtained from tissue sample) is obtained from a subject suspected of having a tumor. In some embodiments, the DNA (e.g., cfDNA or DNA obtained from tissue sample) is obtained from a subject having neoplasia. In some embodiments, the DNA (e.g., cfDNA or DNA obtained from tissue sample) is obtained from a subject suspected of having neoplasia. In some embodiments, the DNA (e.g., cfDNA or DNA obtained from tissue sample) is obtained from a subject in remission from a tumor, cancer, or neoplasia (e.g., following chemotherapy, surgical resection, radiation, or a combination thereof). In any of the foregoing embodiments, the cancer, tumor, or neoplasia or suspected cancer, tumor, or neoplasia may be of the lung, colon, rectum, kidney, breast, prostate, or liver. In some embodiments, the cancer, tumor, or neoplasia or suspected cancer, tumor, or neoplasia is of the lung. In some embodiments, the cancer, tumor, or neoplasia or suspected cancer, tumor, or neoplasia is of the colon or rectum. In some embodiments, the cancer, tumor, or neoplasia or suspected cancer, tumor, or neoplasia is of the breast. In some embodiments, the cancer, tumor, or neoplasia or suspected cancer, tumor, or neoplasia is of the prostate. In any of the foregoing embodiments, the subject may be a human subject.

[00213] In some embodiments, the subject was previously diagnosed with a cancer, e.g., any of the cancers noted above or elsewhere herein. Such a subject may have previously received one or more previous cancer treatments, e.g., surgery, chemotherapy, radiation, and/or immunotherapy. In some embodiments, a sample (e.g., cfDNA or DNA obtained from tissue sample) is obtained from a previously diagnosed and treated subject at one or more preselected time points following the one or more previous cancer treatments.

[00214] The sample (e.g., cfDNA or DNA obtained from tissue sample) obtained from the subject may be sequenced to provide a set of sequence information, which may include sequencing captured DNA molecules of the sequence-variable target region set a greater depth of sequencing than captured DNA molecules of the epigenetic target region set, as described in detail elsewhere herein.

4. Collections of target-specific probes

[00215] In some embodiments, a collection of target-specific probes used in methods disclosed herein, which comprises target-binding probes specific for a sequence-variable target region set and target-binding probes specific for an epigenetic target region set. In some embodiments, the capture yield of the target-binding probes specific for the sequence-variable target region set is higher (e.g., at least 2-fold higher) than the capture yield of the target-binding probes specific for the epigenetic target region set. In some embodiments, the collection of target-specific probes is configured to have a capture yield specific for the sequence-variable target region set higher (e.g., at least 2-fold higher) than its capture yield specific for the epigenetic target region set.

[00216] In some embodiments, the capture yield of the target-binding probes specific for the sequence-variable target region set is at least 1.25-, 1.5-, 1.75-, 2-, 2.25-, 2.5-, 2.75-, 3-, 3.5-, 4-, 4.5-, 5-, 6-, 7-, 8-, 9-, 10-, 11-, 12-, 13-, 14-, or 15-fold higher than the capture yield of the target-binding probes specific for the epigenetic target region set. In some embodiments, the capture yield of the target-binding probes specific for the sequence-variable target region set is 1.25- to 1.5-, 1.5- to 1.75-, 1.75- to 2-, 2- to 2.25-, 2.25- to 2.5-, 2.5- to 2.75-, 2.75- to 3-, 3- to 3.5-, 3.5- to 4-, 4- to 4.5-, 4.5- to 5-, 5- to 5.5-, 5.5- to 6-, 6- to 7-, 7- to 8-, 8- to 9-, 9- to 10-, 10- to 11-, 11- to 12-, 13- to 14-, or 14- to 15-fold higher than the capture yield of the target-binding probes specific for the epigenetic target region set. In some embodiments, the capture yield of the target-binding probes specific for the sequence-variable target region set is at least 5-fold higher than the capture yield of the target-binding probes specific for the epigenetic target region set. In some embodiments, the capture yield of the target-binding probes specific for the sequence-variable target region set is 5- to 10-fold higher than the capture yield of the target-binding probes specific for the epigenetic target region set.

[00217] In some embodiments, the collection of target-specific probes is configured to have a capture yield specific for the sequence-variable target region set at least 1.25-, 1.5-, 1.75-, 2-, 2.25-, 2.5-, 2.75-, 3-, 3.5-, 4-, 4.5-, 5-, 6-, 7-, 8-, 9-, 10-, 11-, 12-, 13-, 14-, or 15-fold higher than its capture yield for the epigenetic target region set. In some embodiments, the collection of target-specific probes is configured to have a capture yield specific for the sequence-variable target region set is 1.25- to 1.5-, 1.5- to 1.75-, 1.75- to 2-, 2- to 2.25-, 2.25- to 2.5-, 2.5- to 2.75-, 2.75- to 3-, 3- to 3.5-, 3.5- to 4-, 4- to 4.5-, 4.5- to 5-, 5- to 5.5-, 5.5- to 6-, 6- to 7-, 7- to 8-, 8- to 9-, 9- to 10-, 10- to 11-, 11- to 12-, 13- to 14-, or 14- to 15-fold higher than its capture yield specific for the

epigenetic target region set. In some embodiments, the collection of target-specific probes is configured to have a capture yield specific for the sequence-variable target region set at least 5-fold higher than the capture yield of the target-binding probes specific for the epigenetic target region set. In some embodiments, the collection of target-specific probes is configured to have a capture yield specific for the sequence-variable target region set 5- to 10-fold higher than the capture yield of the target-binding probes specific for the epigenetic target region set.

[00218] The collection of probes can be configured to provide higher capture yields for the sequence-variable target region set in various ways, including concentration, different lengths and/or chemistries (e.g., that affect affinity), and combinations thereof. Affinity can be modulated by adjusting probe length and/or including nucleotide modifications as discussed below.

[00219] In some embodiments, the target-specific probes specific for the sequence-variable target region set are present at a higher concentration than the target-specific probes specific for the epigenetic target region set. In some embodiments, the concentration of the target-binding probes specific for the sequence-variable target region set is at least 1.25-, 1.5-, 1.75-, 2-, 2.25-, 2.5-, 2.75-, 3-, 3.5-, 4-, 4.5-, 5-, 6-, 7-, 8-, 9-, 10-, 11-, 12-, 13-, 14-, or 15-fold higher than the concentration of the target-binding probes specific for the epigenetic target region set. In some embodiments, the concentration of the target-binding probes specific for the sequence-variable target region set is 1.25- to 1.5-, 1.5- to 1.75-, 1.75- to 2-, 2- to 2.25-, 2.25- to 2.5-, 2.5- to 2.75-, 2.75- to 3-, 3- to 3.5-, 3.5- to 4-, 4- to 4.5-, 4.5- to 5-, 5- to 5.5-, 5.5- to 6-, 6- to 7-, 7- to 8-, 8- to 9-, 9- to 10-, 10- to 11-, 11- to 12-, 13- to 14-, or 14- to 15-fold higher than the concentration of the target-binding probes specific for the epigenetic target region set. In some embodiments, the concentration of the target-binding probes specific for the sequence-variable target region set is at least 2-fold higher than the concentration of the target-binding probes specific for the epigenetic target region set. In such embodiments, concentration may refer to the average mass per volume concentration of individual probes in each set. In some embodiments, the concentration of the target-binding probes specific for the sequence-variable target region set is at least 5-fold higher than the concentration of the target-binding probes specific for the epigenetic target region set. In some embodiments, the concentration of the target-binding probes specific for the sequence-variable target region set is 5- to 10-fold higher than the concentration of the target-binding probes specific for the epigenetic target region set.

[00220] In some embodiments, the target-specific probes specific for the sequence-variable target region set have a higher affinity for their targets than the target-specific probes specific for the epigenetic target region set. Affinity can be modulated in any way known to those skilled in the art, including by using different probe chemistries. For example, certain nucleotide modifications, such as cytosine 5-methylation (in certain sequence contexts), modifications that provide a heteroatom at the 2' sugar position, and LNA nucleotides, can increase stability of double-stranded nucleic acids, indicating that oligonucleotides with such modifications have relatively higher affinity for their complementary sequences. See, e.g., Severin et al., *Nucleic Acids Res.* 39: 8740–8751 (2011); Freier et al., *Nucleic Acids Res.* 25: 4429–4443 (1997); US Patent No. 9,738,894. Also, longer sequence lengths will generally provide increased affinity. Other nucleotide modifications, such as the substitution of the nucleobase hypoxanthine for guanine, reduce affinity by reducing the amount of hydrogen bonding between the oligonucleotide and its complementary sequence. In some embodiments, the target-specific probes specific for the sequence-variable target region set have modifications that increase their affinity for their targets. In some embodiments, alternatively or additionally, the target-specific probes specific for the epigenetic target region set have modifications that decrease their affinity for their targets. In some embodiments, the target-specific probes specific for the sequence-variable target region set have longer average lengths and/or higher average melting temperatures than the target-specific probes specific for the epigenetic target region set. These embodiments may be combined with each other and/or with differences in concentration as discussed above to achieve a desired fold difference in capture yield, such as any fold difference or range thereof described above.

[00221] In some embodiments, the target-specific probes comprise a capture moiety. The capture moiety may be any of the capture moieties described herein, e.g., biotin. In some embodiments, the target-specific probes are linked to a solid support, e.g., covalently or non-covalently such as through the interaction of a binding pair of capture moieties. In some embodiments, the solid support is a bead, such as a magnetic bead.

[00222] In some embodiments, the target-specific probes specific for the sequence-variable target region set and/or the target-specific probes specific for the epigenetic target region set are a bait set as discussed above, e.g., probes comprising capture moieties and sequences selected to tile across a panel of regions, such as genes.

[00223] In some embodiments, the target-specific probes are provided in a single composition. The single composition may be a solution (liquid or frozen). Alternatively, it may be a lyophilizate.

[00224] Alternatively, the target-specific probes may be provided as a plurality of compositions, e.g., comprising a first composition comprising probes specific for the epigenetic target region set and a second composition comprising probes specific for the sequence-variable target region set. These probes may be mixed in appropriate proportions to provide a combined probe composition with any of the foregoing fold differences in concentration and/or capture yield. Alternatively, they may be used in separate capture procedures (e.g., with aliquots of a sample or sequentially with the same sample) to provide first and second compositions comprising captured epigenetic target regions and sequence-variable target regions, respectively.

a. Probes specific for epigenetic target regions

[00225] The probes for the epigenetic target region set may comprise probes specific for one or more types of target regions likely to differentiate DNA from neoplastic (e.g., tumor or cancer) cells from healthy cells, e.g., non-neoplastic circulating cells. Exemplary types of such regions are discussed in detail herein, e.g., in the sections above concerning captured sets. The probes for the epigenetic target region set may also comprise probes for one or more control regions, e.g., as described herein.

[00226] In some embodiments, the probes for the epigenetic target region probe set have a footprint of at least 100 kb, e.g., at least 200 kb, at least 300 kb, or at least 400 kb. In some embodiments, the probes for the epigenetic target region set have a footprint in the range of 100-1000 kb, e.g., 100-200 kb, 200-300 kb, 300-400 kb, 400-500 kb, 500-600 kb, 600-700 kb, 700-800 kb, 800-900 kb, and 900-1,000 kb. In some embodiments, the probes for the epigenetic target region set have a footprint of at least 1000 kb, at least 2000 kb, at least 3000 kb, at least 4000 kb, at least 5000 kb, at least 6000 kb, at least 7000 kb, at least 8000 kb, at least 9000 kb or at least 1 Mb. In some embodiments, the probes for the epigenetic target region set have a footprint in the range of 1 Mb-20 Mb, e.g., 1-1.2 Mb, 1.2-1.4 Mb, 1.4-1.6 Mb, 1.6-1.8 Mb, 1.8-2 Mb, 2-2.25 Mb, 2.25-2.5 Mb, 2.5-2.75 Mb, 2.75-3 Mb, 3-3.25 Mb, 3.25-3.5 Mb, 3.5-3.75 Mb, 3.75-4 Mb, 4-4.25 Mb, 4.25-4.5 Mb, 4.5-4.75 Mb, 4.75-5 Mb, 5-5.5 Mb, 5.5-6 Mb, 6-6.5 Mb, 6.5-7 Mb, 7-7.5 Mb, 7.5-8 Mb, 8-8.5 Mb, 8.5-9 Mb, 9-9.5 Mb, 9.5-10 Mb, 10-12 Mb, 12-14 Mb, 14-16 Mb, 16-18 Mb, and 18-20 Mb.

i. Hypermethylation variable target regions

[00227] In some embodiments, the probes for the epigenetic target region set comprise probes specific for one or more hypermethylation variable target regions. The hypermethylation variable target regions may be any of those set forth above. For example, in some embodiments, the probes specific for hypermethylation variable target regions comprise probes specific for a plurality of loci listed in Table 1, e.g., at least 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 100% of the loci listed in Table 1. In some embodiments, the probes specific for hypermethylation variable target regions comprise probes specific for a plurality of loci listed in Table 2, e.g., at least 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 100% of the loci listed in Table 2. In some embodiments, the probes specific for hypermethylation variable target regions comprise probes specific for a plurality of loci listed in Table 1 or Table 2, e.g., at least 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 100% of the loci listed in Table 1 or Table 2. In some embodiments, for each locus included as a target region, there may be one or more probes with a hybridization site that binds between the transcription start site and the stop codon (the last stop codon for genes that are alternatively spliced) of the gene. In some embodiments, the one or more probes bind within 300 bp of the listed position, e.g., within 200 or 100 bp. In some embodiments, a probe has a hybridization site overlapping the position listed above. In some embodiments, the probes specific for the hypermethylation target regions include probes specific for one, two, three, four, or five subsets of hypermethylation target regions that collectively show hypermethylation in one, two, three, four, or five of breast, colon, kidney, liver, and lung cancers.

ii. Hypomethylation variable target regions

[00228] In some embodiments, the probes for the epigenetic target region set comprise probes specific for one or more hypomethylation variable target regions. The hypomethylation variable target regions may be any of those set forth above. For example, the probes specific for one or more hypomethylation variable target regions may include probes for regions such as repeated elements, e.g., LINE1 elements, Alu elements, centromeric tandem repeats, pericentromeric tandem repeats, and satellite DNA, and intergenic regions that are ordinarily methylated in healthy cells may show reduced methylation in tumor cells.

[00229] In some embodiments, probes specific for hypomethylation variable target regions include probes specific for repeated elements and/or intergenic regions. In some embodiments, probes

specific for repeated elements include probes specific for one, two, three, four, or five of LINE1 elements, Alu elements, centromeric tandem repeats, pericentromeric tandem repeats, and/or satellite DNA.

[00230] Exemplary probes specific for genomic regions that show cancer-associated hypomethylation include probes specific for nucleotides 8403565-8953708 and/or 151104701-151106035 of human chromosome 1. In some embodiments, the probes specific for hypomethylation variable target regions include probes specific for regions overlapping or comprising nucleotides 8403565-8953708 and/or 151104701-151106035 of human chromosome 1.

iii. CTCF binding regions

[00231] In some embodiments, the probes for the epigenetic target region set include probes specific for CTCF binding regions. In some embodiments, the probes specific for CTCF binding regions comprise probes specific for at least 10, 20, 50, 100, 200, or 500 CTCF binding regions, or 10-20, 20-50, 50-100, 100-200, 200-500, or 500-1000 CTCF binding regions, e.g., such as CTCF binding regions described above or in one or more of CTCFBSDB or the Cuddapah et al., Martin et al., or Rhee et al. articles cited above. In some embodiments, the probes for the epigenetic target region set comprise at least 100 bp, at least 200 bp at least 300 bp, at least 400 bp, at least 500 bp, at least 750 bp, or at least 1000 bp upstream and downstream regions of the CTCF binding sites.

iv. Transcription start sites

[00232] In some embodiments, the probes for the epigenetic target region set include probes specific for transcriptional start sites. In some embodiments, the probes specific for transcriptional start sites comprise probes specific for at least 10, 20, 50, 100, 200, or 500 transcriptional start sites, or 10-20, 20-50, 50-100, 100-200, 200-500, or 500-1000 transcriptional start sites, e.g., such as transcriptional start sites listed in DBTSS. In some embodiments, the probes for the epigenetic target region set comprise probes for sequences at least 100 bp, at least 200 bp, at least 300 bp, at least 400 bp, at least 500 bp, at least 750 bp, or at least 1000 bp upstream and downstream of the transcriptional start sites.

v. Focal amplifications

[00233] As noted above, although focal amplifications are somatic mutations, they can be detected by sequencing based on read frequency in a manner analogous to approaches for detecting certain epigenetic changes such as changes in methylation. As such, regions that may show focal amplifications in cancer can be included in the epigenetic target region set, as discussed above. In some embodiments, the probes specific for the epigenetic target region set include probes specific for focal amplifications. In some embodiments, the probes specific for focal amplifications include probes specific for one or more of AR, BRAF, CCND1, CCND2, CCNE1, CDK4, CDK6, EGFR, ERBB2, FGFR1, FGFR2, KIT, KRAS, MET, MYC, PDGFRA, PIK3CA, and RAF1. For example, in some embodiments, the probes specific for focal amplifications include probes specific for one or more of at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, or 18 of the foregoing targets.

vi. Control regions

[00234] It can be useful to include control regions to facilitate data validation. In some embodiments, the probes specific for the epigenetic target region set include probes specific for control methylated regions that are expected to be methylated in essentially all samples. In some embodiments, the probes specific for the epigenetic target region set include probes specific for control hypomethylated regions that are expected to be hypomethylated in essentially all samples.

b. Probes specific for sequence-variable target regions

[00235] The probes for the sequence-variable target region set may comprise probes specific for a plurality of regions known to undergo somatic mutations in cancer. The probes may be specific for any sequence-variable target region set described herein. Exemplary sequence-variable target region sets are discussed in detail herein, e.g., in the sections above concerning captured sets.

[00236] In some embodiments, the sequence-variable target region probe set has a footprint of at least 10 kb, e.g., at least 20 kb, at least 30 kb, or at least 40 kb. In some embodiments, the sequence-variable target region probe set has a footprint in the range of 10-100 kb, e.g., 10-20 kb, 20-30 kb, 30-40 kb, 40-50 kb, 50-60 kb, 60-70 kb, 70-80 kb, 80-90 kb, and 90-100 kb. In some embodiments, the sequence-variable target region probe set has a footprint of at least 10 kb, at least 20 kb, at least 30 kb, at least 40 kb, at least 50 kb, at least 75 kb, at least 100 kb, at least 200 kb, at least 300 kb, or at least 400 kb. In some embodiments, the sequence-variable target region probe set has a footprint in the range of 100-1000 kb, e.g., 100-200 kb, 200-300 kb, 300-400 kb, 400-500 kb, 500-

600 kb, 600-700 kb, 700-800 kb, 800-900 kb, and 900-1000 kb. In some embodiments, the sequence-variable target region probe set has a footprint of at least 1000 kb, at least 2000 kb, at least 3000 kb, at least 4000 kb, at least 5000 kb, at least 6000 kb, at least 7000 kb, at least 8000 kb, at least 9000 kb or at least 1 Mb. In some embodiments, the sequence-variable target region probe set has a footprint in the range of 1 Mb-10 Mb, e.g., 1-1.2 Mb, 1.2-1.4 Mb, 1.4-1.6 Mb, 1.6-1.8 Mb, 1.8-2 Mb, 2-2.25 Mb, 2.25-2.5 Mb, 2.5-2.75 Mb, 2.75-3 Mb, 3-3.25 Mb, 3.25-3.5 Mb, 3.5-3.75 Mb, 3.75-4 Mb, 4-4.25 Mb, 4.25-4.5 Mb, 4.5-4.75 Mb, 4.75-5 Mb, 5-5.5 Mb, 5.5-6 Mb, 6-6.5 Mb, 6.5-7 Mb, 7-7.5 Mb, 7.5-8 Mb, 8-8.5 Mb, 8.5-9 Mb, 9-9.5 Mb, and 9.5-10 Mb.

[00237] In some embodiments, probes specific for the sequence-variable target region set comprise probes specific for at least a portion of 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 70 of the genes of Table 3. In some embodiments, probes specific for the sequence-variable target region set comprise probes specific for the 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 70 of the SNVs of Table 3. In some embodiments, probes specific for the sequence-variable target region set comprise probes specific for at least 1, at least 2, at least 3, at least 4, at least 5, or 6 of the fusions of Table 3. In some embodiments, probes specific for the sequence-variable target region set comprise probes specific for at least a portion of at least 1, at least 2, or 3 of the indels of Table 3. In some embodiments, probes specific for the sequence-variable target region set comprise probes specific for at least a portion of 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, at least 70, or 73 of the genes of Table 4. In some embodiments, probes specific for the sequence-variable target region set comprise probes specific for 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, at least 70, or 73 of the SNVs of Table 4. In some embodiments, probes specific for the sequence-variable target region set comprise probes specific for at least 1, at least 2, at least 3, at least 4, at least 5, or 6 of the fusions of Table 4. In some embodiments, probes specific for the sequence-variable target region set comprise probes specific for at least a portion of at least 1, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, or 18 of the indels of Table 4. In some embodiments, probes specific for the sequence-variable target region set comprise probes specific

for at least a portion of at least 1, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, or at least 20 of the genes of Table 5.

[00238] In some embodiments, the probes specific for the sequence-variable target region set comprise probes specific for target regions from at least 10, 20, 30, or 35 cancer-related genes, such as AKT1, ALK, BRAF, CCND1, CDK2A, CTNNB1, EGFR, ERBB2, ESR1, FGFR1, FGFR2, FGFR3, FOXL2, GATA3, GNA11, GNAQ, GNAS, HRAS, IDH1, IDH2, KIT, KRAS, MED12, MET, MYC, NFE2L2, NRAS, PDGFRA, PIK3CA, PPP2R1A, PTEN, RET, STK11, TP53, and U2AF1.

c. Compositions of probes

[00239] In some embodiments, a single composition is used, comprising probes for the sequence-variable target region set and probes for the epigenetic target region set. The probes may be provided in such a composition at any concentration ratio described herein.

[00240] In some embodiments, a first composition comprising probes for the epigenetic target region set and a second composition comprising probes for the sequence-variable target region set are provided. The ratio of the concentration of the probes in the first composition to the concentration of the probes in the second composition may be any of the ratios described herein.

5. Compositions comprising captured cfDNA

[00241] In some embodiments, compositions comprising captured cfDNA are generated and/or used in methods disclosed herein. The captured cfDNA may have any of the features described herein concerning captured sets, including, e.g., a greater concentration of the DNA corresponding to the sequence-variable target region set (normalized for footprint size as discussed above) than of the DNA corresponding to the epigenetic target region set. In some embodiments, the cfDNA of the captured set comprises sequence tags, which may be added to the cfDNA as described herein. In general, the inclusion of sequence tags results in the cfDNA molecules differing from their naturally occurring, untagged form.

[00242] Such compositions may further comprise a probe set described herein or sequencing primers, each of which may differ from naturally occurring nucleic acid molecules. For example, a probe set described herein may comprise a capture moiety, and sequencing primers may comprise a non-naturally occurring label.

6. Exemplary method for molecular tag identification of MBD-bead partitioned libraries

[00243] An exemplary method for molecular tag identification of MBD-bead partitioned libraries through NGS is as follows:

- i) Physical partitioning of an extracted DNA sample (e.g., extracted blood plasma DNA from a human sample, which has optionally been subjected to target capture as described herein) using a methyl-binding domain protein-bead purification kit, saving all elutions from process for downstream processing.
- ii) Parallel application of differential molecular tags and NGS-enabling adapter sequences to each partition. For example, the hypermethylated, residual methylation ('wash'), and hypomethylated partitions are ligated with NGS- adapters with molecular tags.
- iii) Re-combining all molecular tagged partitions, and subsequent amplification using adapter-specific DNA primer sequences.
- iv) Capture/hybridization of re-combined and amplified total library, targeting genomic regions of interest (e.g., cancer-specific genetic variants and differentially methylated regions).
- v) Re-amplification of the captured DNA library, appending a sample tag. Different samples are pooled and assayed in multiplex on an NGS instrument.
- vi) Bioinformatics analysis of NGS data, with the molecular tags being used to identify unique molecules, as well deconvolution of the sample into molecules that were differentially MBD-partitioned. This analysis can yield information on relative 5-methylcytosine for genomic regions, concurrent with standard genetic sequencing/variant detection. As discussed in detail elsewhere herein, the analysis may comprise determining a sequence from the first and second sets of sequence reads, wherein (i) calling a C to T or G to A transition mutation relative to a reference sequence based on reads or sequences of molecules of the first set requires observation of the transition mutation in a greater number of reads than calling a C to T or G to A transition mutation relative to the reference sequence based on reads or sequences of molecules of the second set; or (ii) C to T or G to A transition mutations are not called relative to a reference sequence based on reads or

sequences of molecules of the first set.

[00244] The exemplary method set forth above may further comprise any compatible feature of methods according to this disclosure set forth elsewhere herein.

7. Exemplary workflows

[00245] Exemplary workflows for partitioning and library preparation are provided herein. In some embodiments, some or all features of the partitioning and library preparation workflows may be used in combination. The exemplary workflows set forth above may further comprise any compatible feature of methods according to this disclosure set forth elsewhere herein.

a. Partitioning

[00246] In some embodiments, sample DNA (for e.g., between 1 and 300 ng) is mixed with an appropriate amount of methyl binding domain (MBD) buffer (the amount of MBD buffer depends on the amount of DNA used) and magnetic beads conjugated with MBD proteins and incubated overnight. Methylated DNA (hypermethylated DNA) binds the MBD protein on the magnetic beads during this incubation. Non-methylated (hypomethylated DNA) or less methylated DNA (intermediately methylated) is washed away from the beads with buffers containing increasing concentrations of salt. For example, one, two, or more fractions containing non-methylated, hypomethylated, and/or intermediately methylated DNA may be obtained from such washes. Finally, a high salt buffer is used to elute the heavily methylated DNA (hypermethylated DNA) from the MBD protein. In some embodiments, these washes result in three partitions (hypomethylated partition, intermediately methylated fraction and hypermethylated partition) of DNA having increasing levels of methylation.

[00247] In some embodiments, the three partitions of DNA are desalted and concentrated in preparation for the enzymatic steps of library preparation.

b. Library preparation

[00248] In some embodiments (e.g., after concentrating the DNA in the partitions), the partitioned DNA is made ligatable, e.g., by extending the end overhangs of the DNA molecules are extended, and adding adenosine residues to the 3' ends of fragments and phosphorylating the 5' end of each DNA fragment. DNA ligase and adapters are added to ligate each partitioned DNA molecule with an adapter on each end. These adapters contain partition tags (e.g., non-random, non-unique barcodes) that are distinguishable from the partition tags in the adapters used in the other partitions.

After ligation, the three partitions are pooled together and are amplified (e.g., by PCR, such as with primers specific for the adapters).

[00249] Following PCR, amplified DNA may be cleaned and concentrated prior to capture. The amplified DNA is contacted with a collection of probes described herein (which may be, e.g., biotinylated RNA probes or ssDNA probes or dsDNA probes) that target specific regions of interest. The mixture is incubated, e.g., overnight, e.g., in a salt buffer. The probes are captured (e.g., using streptavidin magnetic beads) and separated from the amplified DNA that was not captured, such as by a series of salt washes, thereby providing a captured set of DNA. After capture, the DNA of the captured set is amplified by PCR. In some embodiments, the PCR primers contain a sample tag, thereby incorporating the sample tag into the DNA molecules. In some embodiments, DNA from different samples is pooled together and then multiplex sequenced, e.g., using an Illumina NovaSeq sequencer.

III. GENERAL FEATURES OF THE METHODS

1. Samples

[00250] A sample can be any biological sample isolated from a subject. A sample can be a bodily sample. Samples can include body tissues, such as known or suspected solid tumors, whole blood, platelets, serum, plasma, stool, red blood cells, white blood cells or leucocytes, endothelial cells, tissue biopsies, cerebrospinal fluid synovial fluid, lymphatic fluid, ascites fluid, interstitial or extracellular fluid, the fluid in spaces between cells, including gingival crevicular fluid, bone marrow, pleural effusions, cerebrospinal fluid, saliva, mucous, sputum, semen, sweat, urine. Samples are preferably body fluids, particularly blood and fractions thereof, and urine. A sample can be in the form originally isolated from a subject or can have been subjected to further processing to remove or add components, such as cells, or enrich for one component relative to another. Thus, a preferred body fluid for analysis is plasma or serum containing cell-free nucleic acids. A sample can be isolated or obtained from a subject and transported to a site of sample analysis. The sample may be preserved and shipped at a desirable temperature, e.g., room temperature, 4°C, -20°C, and/or -80°C. A sample can be isolated or obtained from a subject at the site of the sample analysis. The subject can be a human, a mammal, an animal, a companion animal, a service animal, or a pet. The subject may have a cancer. The subject may not have cancer

or a detectable cancer symptom. The subject may have been treated with one or more cancer therapy, e.g., any one or more of chemotherapies, antibodies, vaccines or biologics. The subject may be in remission. The subject may or may not be diagnosed of being susceptible to cancer or any cancer-associated genetic mutations/disorders.

[00251] The volume of plasma can depend on the desired read depth for sequenced regions. Exemplary volumes are 0.4-40 ml, 5-20 ml, 10-20 ml. For examples, the volume can be 0.5 mL, 1 mL, 5 mL 10 mL, 20 mL, 30 mL, or 40 mL. A volume of sampled plasma may be 5 to 20 mL.

[00252] In some embodiments, the sample can be a DNA sample obtained from a tissue. In such embodiments, the DNA obtained from the tissue sample can be fragmented by enzymatic (e.g. fragmentase) or mechanical means (e.g. shearing by sonication).

[00253] A sample can comprise various amount of nucleic acid that contains genome equivalents. For example, a sample of about 30 ng DNA can contain about 10,000 (10^4) haploid human genome equivalents and, in the case of cfDNA, about 200 billion (2×10^{11}) individual polynucleotide molecules. Similarly, a sample of about 100 ng of DNA can contain about 30,000 haploid human genome equivalents and, in the case of cfDNA, about 600 billion individual molecules.

[00254] A sample can comprise nucleic acids from different sources, e.g., from cells and cell-free of the same subject, from cells and cell-free of different subjects. A sample can comprise nucleic acids carrying mutations. For example, a sample can comprise DNA carrying germline mutations and/or somatic mutations. Germline mutations refer to mutations existing in germline DNA of a subject. Somatic mutations refer to mutations originating in somatic cells of a subject, e.g., cancer cells. A sample can comprise DNA carrying cancer-associated mutations (e.g., cancer-associated somatic mutations). A sample can comprise an epigenetic variant (i.e. a chemical or protein modification), wherein the epigenetic variant associated with the presence of a genetic variant such as a cancer-associated mutation. In some embodiments, the sample comprises an epigenetic variant associated with the presence of a genetic variant, wherein the sample does not comprise the genetic variant.

[00255] Exemplary amounts of cell-free nucleic acids in a sample before amplification range from about 1 fg to about 1 μ g, e.g., 1 pg to 200 ng, 1 ng to 100 ng, 10 ng to 1000 ng. For example, the amount can be up to about 600 ng, up to about 500 ng, up to about 400 ng, up to about 300 ng, up to about 200 ng, up to about 100 ng, up to about 50 ng, or up to about 20 ng of cell-free nucleic acid molecules. The amount can be at least 1 fg, at least 10 fg, at least 100 fg, at least 1 pg, at least

10 pg, at least 100 pg, at least 1 ng, at least 10 ng, at least 100 ng, at least 150 ng, or at least 200 ng of cell-free nucleic acid molecules. The amount can be up to 1 femtogram (fg), 10 fg, 100 fg, 1 picogram (pg), 10 pg, 100 pg, 1 ng, 10 ng, 100 ng, 150 ng, 200 ng, 250 ng or 300 ng of cell-free nucleic acid molecules. The method can comprise obtaining 1 femtogram (fg) to 200 ng. In some embodiments, the amount of DNA used can be between 1 fg and 1 µg.

[00256] Cell-free nucleic acids are nucleic acids not contained within or otherwise bound to a cell or in other words nucleic acids remaining in a sample after removing intact cells. Cell-free nucleic acids include DNA, RNA, and hybrids thereof, including genomic DNA, mitochondrial DNA, siRNA, miRNA, circulating RNA (cRNA), tRNA, rRNA, small nucleolar RNA (snoRNA), Piwi-interacting RNA (piRNA), long non-coding RNA (long ncRNA), or fragments of any of these. Cell-free nucleic acids can be double-stranded, single-stranded, or a hybrid thereof. A cell-free nucleic acid can be released into bodily fluid through secretion or cell death processes, e.g., cellular necrosis and apoptosis. Some cell-free nucleic acids are released into bodily fluid from cancer cells e.g., circulating tumor DNA, (ctDNA). Others are released from healthy cells. In some embodiments, cfDNA is cell-free fetal DNA (cffDNA) In some embodiments, cell free nucleic acids are produced by tumor cells. In some embodiments, cell free nucleic acids are produced by a mixture of tumor cells and non-tumor cells.

[00257] Cell-free nucleic acids have an exemplary size distribution of about 100-500 nucleotides, with molecules of 110 to about 230 nucleotides representing about 90% of molecules, with a mode of about 168 nucleotides and a second minor peak in a range between 240 to 440 nucleotides.

[00258] In some embodiments, the DNA in a sample consists essentially of cell-free DNA. This means that all or nearly all of the DNA in the sample is cfDNA, e.g., at least 90% of the DNA by weight or mole fraction. In some embodiments, at least 95%, 97%, 98%, 99%, 99.5%, or 99.9% of the DNA by weight or mole fraction in a sample is cfDNA. In some embodiments, the DNA in a sample consists of cell-free DNA.

[00259] Cell-free nucleic acids can be isolated from bodily fluids through a fractionation or partitioning step in which cell-free nucleic acids, as found in solution, are separated from intact cells and other non-soluble components of the bodily fluid. Partitioning may include techniques such as centrifugation or filtration. Alternatively, cells in bodily fluids can be lysed and cell-free and cellular nucleic acids processed together. Generally, after addition of buffers and wash steps, nucleic acids can be precipitated with an alcohol. Further clean up steps may be used such as silica-

based columns to remove contaminants or salts. Non-specific bulk carrier nucleic acids, such as C 1 DNA, DNA or protein for bisulfite sequencing, hybridization, and/or ligation, may be added throughout the reaction to optimize certain aspects of the procedure such as yield.

[00260] After such processing, samples can include various forms of nucleic acid including double stranded DNA, single stranded DNA and single stranded RNA. In some embodiments, single stranded DNA and RNA can be converted to double stranded forms so they are included in subsequent processing and analysis steps.

[00261] Double-stranded DNA molecules in a sample and single stranded nucleic acid molecules converted to double stranded DNA molecules can be linked to adapters at either one end or both ends. Typically, double stranded molecules are blunt ended by treatment with a polymerase with a 5'-3' polymerase and a 3' '-5' exonuclease (or proof-reading function), in the presence of all four standard nucleotides. Klenow large fragment and T4 polymerase are examples of suitable polymerase. The blunt ended DNA molecules can be ligated with at least partially double stranded adapter (e.g., a Y shaped or bell-shaped adapter). Alternatively, complementary nucleotides can be added to blunt ends of sample nucleic acids and adapters to facilitate ligation. Contemplated herein are both blunt end ligation and sticky end ligation. In blunt end ligation, both the nucleic acid molecules and the adapter tags have blunt ends. In sticky-end ligation, typically, the nucleic acid molecules bear an "A" overhang and the adapters bear a "T" overhang.

2. Tags

[00262] Tags comprising barcodes can be incorporated into or otherwise joined to adapters. Tags can be incorporated by ligation, overlap extension PCR among other methods.

a. Molecular tagging strategies

[00263] Molecular tagging refers to a tagging practice that allows one to differentiate molecules from which sequence reads originated. Tagging strategies can be divided into unique tagging and non-unique tagging strategies. In unique tagging, all or substantially all of the molecules in a sample bear a different tag, so that reads can be assigned to original molecules based on tag information alone. Tags used in such methods are sometimes referred to as "unique tags". In non-unique tagging, different molecules in the same sample can bear the same tag, so that other information in addition to tag information is used to assign a sequence read to an original molecule. Such information may include start and stop coordinate, coordinate to which the molecule maps,

start or stop coordinate alone, etc. Tags used in such methods are sometimes referred to as “non-unique tags”. Accordingly, it is not necessary to uniquely tag every molecule in a sample. It suffices to uniquely tag molecules falling within an identifiable class within a sample. Thus, molecules in different identifiable families can bear the same tag without loss of information about the identity of the tagged molecule.

[00264] In certain embodiments of non-unique tagging, the number of different tags used can be sufficient that there is a very high likelihood (e.g., at least 99%, at least 99.9%, at least 99.99% or at least 99.999%) that all molecules of a particular group bear a different tag. It is to be noted that when barcodes are used as tags, and when barcodes are attached, e.g., randomly, to both ends of a molecule, the combination of barcodes, together, can constitute a tag. This number, in turn, is a function of the number of molecules falling into the class. For example, the class may be all molecules mapping to the same start-stop position on a reference genome. The class may be all molecules mapping across a particular genetic locus, e.g., a particular base or a particular region (e.g., up to 100 bases or a gene or an exon of a gene). In certain embodiments, the number of different tags used to uniquely identify a number of molecules, z , in a class can be between any of $2*z$, $3*z$, $4*z$, $5*z$, $6*z$, $7*z$, $8*z$, $9*z$, $10*z$, $11*z$, $12*z$, $13*z$, $14*z$, $15*z$, $16*z$, $17*z$, $18*z$, $19*z$, $20*z$ or $100*z$ (e.g., lower limit) and any of $100,000*z$, $10,000*z$, $1000*z$ or $100*z$ (e.g., upper limit).

[00265] For example, in a sample of about 3 ng to 30 ng of human cell free DNA, one expects around 10^3 - 10^4 molecules to map to a particular nucleotide coordinate, and between about 3 and 10 molecules having any start coordinate to share the same stop coordinate. Accordingly, about 50 to about 50,000 different tags (e.g., between about 6 and 220 barcode combinations) can suffice to uniquely tag all such molecules. To uniquely tag all 10^3 - 10^4 molecules mapping across a nucleotide coordinate, about 1 million to about 20 million different tags would be required.

[00266] Generally, assignment of unique or non-unique tags barcodes in reactions follows methods and systems described by US patent applications 20010053519, 20030152490, 20110160078, and U.S. Pat. No. 6,582,908 and U.S. Pat. No. 7,537,898 and US Pat. No. 9,598,731. Tags can be linked to sample nucleic acids randomly or non-randomly.

[00267] In some embodiments, the tagged nucleic acids are sequenced after loading into a microwell plate. The microwell plate can have 96, 384, or 1536 microwells. In some cases, they are introduced at an expected ratio of unique tags to microwells. For example, the unique tags may

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

[00268] A preferred format uses 20-50 different tags (e.g., barcodes) ligated to both ends of target nucleic acids. For example, 35 different tags (e.g., barcodes) ligated to both ends of target molecules creating 35 x 35 permutations, which equals 1225 tag combinations for 35 tags. Such numbers of tags are sufficient so that different molecules having the same start and stop points have a high probability (e.g., at least 94%, 99.5%, 99.99%, 99.999%) of receiving different combinations of tags. Other barcode combinations include any number between 10 and 500, e.g., about 15x15, about 35x35, about 75x75, about 100x100, about 250x250, about 500x500.

[00269] In some cases, unique tags may be predetermined or random or semi-random sequence oligonucleotides. In other cases, a plurality of barcodes may be used such that barcodes are not necessarily unique to one another in the plurality. In this example, barcodes may be ligated to individual molecules such that the combination of the barcode and the sequence it may be ligated to creates a unique sequence that may be individually tracked. As described herein, detection of non-unique barcodes in combination with sequence data of beginning (start) and end (stop) portions of sequence reads may allow assignment of a unique identity to a particular molecule. The length or number of base pairs, of an individual sequence read may also be used to assign a unique identity to such a molecule. As described herein, fragments from a single strand of nucleic acid having been assigned a unique identity, may thereby permit subsequent identification of fragments from the parent strand.

3. Amplification

[00270] Sample nucleic acids flanked by adapters can be amplified by PCR and other amplification methods. Amplification is typically primed by primers binding to primer binding sites in adapters flanking a DNA molecule to be amplified. Amplification methods can involve cycles of

denaturation, annealing and extension, resulting from thermocycling or can be isothermal as in transcription-mediated amplification. Other amplification methods include the ligase chain reaction, strand displacement amplification, nucleic acid sequence-based amplification, and self-sustained sequence-based replication.

[00271] Preferably, the present methods perform dsDNA ligations with T-tailed and C-tailed adapters, which result in amplification of at least 50, 60, 70 or 80% of double stranded nucleic acids before linking to adapters. Preferably the present methods increase the amount or number of amplified molecules relative to control methods performed with T-tailed adapters alone by at least 10, 15 or 20%.

4. Bait sets; Capture moieties; Enrichment

[00272] As discussed above, nucleic acids in a sample can be subject to a capture step, in which molecules having target sequences are captured for subsequent analysis. Target capture can involve use of a bait set comprising oligonucleotide baits labeled with a capture moiety, such as biotin or the other examples noted below. The probes can have sequences selected to tile across a panel of regions, such as genes. In some embodiments, a bait set can have higher and lower capture yields for sets of target regions such as those of the sequence-variable target region set and the epigenetic target region set, respectively, as discussed elsewhere herein. In some embodiments, the baits (i.e., probes) can be RNA, ssDNA or dsDNA. The bait sets are combined with a sample under conditions that allow hybridization of the target molecules with the baits. Then, captured molecules are isolated using the capture moiety. For example, a biotin capture moiety by bead-based streptavidin. Such methods are further described in, for example, U.S. patent 9,850,523, issuing December 26, 2017, which is incorporated herein by reference.

[00273] Capture moieties include, without limitation, biotin, avidin, streptavidin, a nucleic acid comprising a particular nucleotide sequence, a hapten recognized by an antibody, and magnetically attractable particles. The extraction moiety can be a member of a binding pair, such as biotin/streptavidin or hapten/antibody. In some embodiments, a capture moiety that is attached to an analyte is captured by its binding pair which is attached to an isolatable moiety, such as a magnetically attractable particle or a large particle that can be sedimented through centrifugation. The capture moiety can be any type of molecule that allows affinity separation of nucleic acids bearing the capture moiety from nucleic acids lacking the capture moiety. Exemplary capture

moieties are biotin which allows affinity separation by binding to streptavidin linked or linkable to a solid phase or an oligonucleotide, which allows affinity separation through binding to a complementary oligonucleotide linked or linkable to a solid phase.

5. Sequencing

[00274] Sample nucleic acids, optionally flanked by adapters, with or without prior amplification are generally subjected to sequencing. Sequencing methods or commercially available formats that are optionally utilized include, for example, Sanger sequencing, high-throughput sequencing, pyrosequencing, sequencing-by-synthesis, single-molecule sequencing, nanopore-based sequencing, semiconductor sequencing, sequencing-by-ligation, sequencing-by-hybridization, RNA-Seq (Illumina), Digital Gene Expression (Helicos), next generation sequencing (NGS), Single Molecule Sequencing by Synthesis (SMSS) (Helicos), massively-parallel sequencing, Clonal Single Molecule Array (Solexa), shotgun sequencing, Ion Torrent, Oxford Nanopore, Roche Genia, Maxim-Gilbert sequencing, primer walking, sequencing using PacBio, SOLiD, Ion Torrent, or Nanopore platforms. The principles described herein, including more stringently calling C to T or G to A transition mutations detected from reads or sequences of molecules from a hypermethylated partition, can be applied by those skilled in the art to sequencing approaches that directly detect methylation, e.g., sequencing using Oxford Nanopore or PacBio. Sequencing reactions can be performed in a variety of sample processing units, which may include multiple lanes, multiple channels, multiple wells, or other means of processing multiple sample sets substantially simultaneously. Sample processing units can also include multiple sample chambers to enable the processing of multiple runs simultaneously.

[00275] The sequencing reactions can be performed on one or more nucleic acid fragment types or regions containing markers of cancer or of other diseases. The sequencing reactions can also be performed on any nucleic acid fragment present in the sample. The sequence reactions may be performed on at least about 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 99%, 99.9%, or 100% of the genome. In other cases, sequence reactions may be performed on less than about 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 99%, 99.9%, or 100% of the genome.

[00276] Simultaneous sequencing reactions may be performed using multiplex sequencing techniques. In some embodiments, cell-free polynucleotides are sequenced with at least about

1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, or 100,000 sequencing reactions. In other embodiments, cell-free polynucleotides are sequenced with less than about 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, or 100,000 sequencing reactions. Sequencing reactions are typically performed sequentially or simultaneously. Subsequent data analysis is generally performed on all or part of the sequencing reactions. In some embodiments, data analysis is performed on at least about 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, or 100,000 sequencing reactions. In other embodiments, data analysis may be performed on less than about 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, or 100,000 sequencing reactions. An example of a read depth is from about 1000 to about 50000 reads per locus (e.g., base position).

a. Differential depth of sequencing

[00277] In some embodiments, nucleic acids corresponding to the sequence-variable target region set are sequenced to a greater depth of sequencing than nucleic acids corresponding to the epigenetic target region set. For example, the depth of sequencing for nucleic acids corresponding to the sequence variant target region set may be at least 1.25-, 1.5-, 1.75-, 2-, 2.25-, 2.5-, 2.75-, 3-, 3.5-, 4-, 4.5-, 5-, 6-, 7-, 8-, 9-, 10-, 11-, 12-, 13-, 14-, or 15-fold greater, or 1.25- to 1.5-, 1.5- to 1.75-, 1.75- to 2-, 2- to 2.25-, 2.25- to 2.5-, 2.5- to 2.75-, 2.75- to 3-, 3- to 3.5-, 3.5- to 4-, 4- to 4.5-, 4.5- to 5-, 5- to 5.5-, 5.5- to 6-, 6- to 7-, 7- to 8-, 8- to 9-, 9- to 10-, 10- to 11-, 11- to 12-, 13- to 14-, 14- to 15-fold, or 15- to 100-fold greater, than the depth of sequencing for nucleic acids corresponding to the epigenetic target region set. In some embodiments, said depth of sequencing is at least 2-fold greater. In some embodiments, said depth of sequencing is at least 5-fold greater. In some embodiments, said depth of sequencing is at least 10-fold greater. In some embodiments, said depth of sequencing is 4- to 10-fold greater. In some embodiments, said depth of sequencing is 4- to 100-fold greater. Each of these embodiments refer to the extent to which nucleic acids corresponding to the sequence-variable target region set are sequenced to a greater depth of sequencing than nucleic acids corresponding to the epigenetic target region set.

[00278] In some embodiments, the captured cfDNA corresponding to the sequence-variable target region set and the captured cfDNA corresponding to the epigenetic target region set are sequenced concurrently, e.g., in the same sequencing cell (such as the flow cell of an Illumina sequencer) and/or in the same composition, which may be a pooled composition resulting from recombining separately captured sets or a composition obtained by capturing the cfDNA corresponding to the

sequence-variable target region set and the captured cfDNA corresponding to the epigenetic target region set in the same vessel.

b. Preparation for sequencing

[00279] In some embodiments, a nucleic acid population is prepared for sequencing by enzymatically forming blunt-ends on double-stranded nucleic acids with single-stranded overhangs at one or both ends. In these embodiments, the population is typically treated with an enzyme having a 5'-3' DNA polymerase activity and a 3'-5' exonuclease activity in the presence of the nucleotides (e.g., A, C, G, and T or U). Examples of enzymes or catalytic fragments thereof that may be optionally used include Klenow large fragment and T4 polymerase. At 5' overhangs, the enzyme typically extends the recessed 3' end on the opposing strand until it is flush with the 5' end to produce a blunt end. At 3' overhangs, the enzyme generally digests from the 3' end up to and sometimes beyond the 5' end of the opposing strand. If this digestion proceeds beyond the 5' end of the opposing strand, the gap can be filled in by an enzyme having the same polymerase activity that is used for 5' overhangs. The formation of blunt ends on double-stranded nucleic acids facilitates, for example, the attachment of adapters and subsequent amplification.

[00280] In some embodiments, nucleic acid populations are subjected to additional processing, such as the conversion of single-stranded nucleic acids to double-stranded nucleic acids and/or conversion of RNA to DNA (e.g., complementary DNA or cDNA). These forms of nucleic acid are also optionally linked to adapters and amplified.

[00281] With or without prior amplification, nucleic acids subject to the process of forming blunt-ends described above, and optionally other nucleic acids in a sample, can be sequenced to produce sequenced nucleic acids. A sequenced nucleic acid can refer either to the sequence of a nucleic acid (e.g., sequence information) or a nucleic acid whose sequence has been determined. Sequencing can be performed so as to provide sequence data of individual nucleic acid molecules in a sample either directly or indirectly from a consensus sequence of amplification products of an individual nucleic acid molecule in the sample.

[00282] In some embodiments, double-stranded nucleic acids with single-stranded overhangs in a sample after blunt-end formation are linked at both ends to adapters including barcodes, and the sequencing determines nucleic acid sequences as well as in-line barcodes introduced by the adapters. The blunt-end DNA molecules are optionally ligated to a blunt end of an at least partially double-stranded adapter (e.g., a Y-shaped or bell-shaped adapter). Alternatively, blunt ends of

sample nucleic acids and adapters can be tailed with complementary nucleotides to facilitate ligation (for e.g., sticky-end ligation).

[00283] The nucleic acid sample is typically contacted with a sufficient number of adapters that there is a low probability (e.g., less than about 1 or 0.1 %) that any two copies of the same nucleic acid receive the same combination of adapter barcodes from the adapters linked at both ends. The use of adapters in this manner may permit identification of families of nucleic acid sequences with the same start and stop points on a reference nucleic acid and linked to the same combination of barcodes. Such a family may represent sequences of amplification products of a nucleic acid in the sample before amplification. The sequences of family members can be compiled to derive consensus nucleotide(s) or a complete consensus sequence for a nucleic acid molecule in the original sample, as modified by blunt-end formation and adapter attachment. In other words, the nucleotide occupying a specified position of a nucleic acid in the sample can be determined to be the consensus of nucleotides occupying that corresponding position in family member sequences. Families can include sequences of one or both strands of a double-stranded nucleic acid. If members of a family include sequences of both strands from a double-stranded nucleic acid, sequences of one strand may be converted to their complements for purposes of compiling sequences to derive consensus nucleotide(s) or sequences. Some families include only a single member sequence. In this case, this sequence can be taken as the sequence of a nucleic acid in the sample before amplification. Alternatively, families with only a single member sequence can be eliminated from subsequent analysis.

[00284] Nucleotide variations (e.g., SNVs or indels) in sequenced nucleic acids can be determined by comparing sequenced nucleic acids with a reference sequence. The reference sequence is often a known sequence, e.g., a known whole or partial genome sequence from a subject (e.g., a whole genome sequence of a human subject). The reference sequence can be, for example, hG19 or hG38. The sequenced nucleic acids can represent sequences determined directly for a nucleic acid in a sample, or a consensus of sequences of amplification products of such a nucleic acid, as described above. A comparison can be performed at one or more designated positions on a reference sequence. A subset of sequenced nucleic acids can be identified including a position corresponding with a designated position of the reference sequence when the respective sequences are maximally aligned. Within such a subset it can be determined which, if any, sequenced nucleic acids include a nucleotide variation at the designated position, and optionally which if any, include

a reference nucleotide (e.g., same as in the reference sequence). If the number of sequenced nucleic acids in the subset including a nucleotide variant exceeding a selected threshold, then a variant nucleotide can be called at the designated position. The threshold can be a simple number, such as at least 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 sequenced nucleic acids within the subset including the nucleotide variant or it can be a ratio, such as at least 0.5, 1, 2, 3, 4, 5, 10, 15, or 20, of sequenced nucleic acids within the subset that include the nucleotide variant, among other possibilities. The comparison can be repeated for any designated position of interest in the reference sequence. Sometimes a comparison can be performed for designated positions occupying at least about 20, 100, 200, or 300 contiguous positions on a reference sequence, e.g., about 20-500, or about 50-300 contiguous positions.

[00285] Additional details regarding nucleic acid sequencing, including the formats and applications described herein, are also provided in, for example, Levy et al., *Annual Review of Genomics and Human Genetics*, 17: 95-115 (2016), Liu et al., *J. of Biomedicine and Biotechnology*, Volume 2012, Article ID 251364:1-11 (2012), Voelkerding et al., *Clinical Chem.*, 55: 641-658 (2009), MacLean et al., *Nature Rev. Microbiol.*, 7: 287-296 (2009), Astier et al., *J Am Chem Soc.*, 128(5):1705-10 (2006), U.S. Pat. No. 6,210,891, U.S. Pat. No. 6,258,568, U.S. Pat. No. 6,833,246, U.S. Pat. No. 7,115,400, U.S. Pat. No. 6,969,488, U.S. Pat. No. 5,912,148, U.S. Pat. No. 6,130,073, U.S. Pat. No. 7,169,560, U.S. Pat. No. 7,282,337, U.S. Pat. No. 7,482,120, U.S. Pat. No. 7,501,245, U.S. Pat. No. 6,818,395, U.S. Pat. No. 6,911,345, U.S. Pat. No. 7,501,245, U.S. Pat. No. 7,329,492, U.S. Pat. No. 7,170,050, U.S. Pat. No. 7,302,146, U.S. Pat. No. 7,313,308, and U.S. Pat. No. 7,476,503, each of which is hereby incorporated by reference in its entirety.

6. Analysis

[00286] Sequencing may generate a plurality of sequence reads or reads. Sequence reads or reads may include sequences of nucleotide data less than about 150 bases in length, or less than about 90 bases in length. In some embodiments, reads are between about 80 bases and about 90 bases, e.g., about 85 bases in length. In some embodiments, methods of the present disclosure are applied to very short reads, e.g., less than about 50 bases or about 30 bases in length. Sequence read data can include the sequence data as well as meta information. Sequence read data can be stored in any suitable file format including, for example, VCF files, FASTA files, or FASTQ files.

[00287] FASTA may refer to a computer program for searching sequence databases, and the name FASTA may also refer to a standard file format. FASTA is described by, for example, Pearson & Lipman, 1988, Improved tools for biological sequence comparison, PNAS 85:2444-2448, which is hereby incorporated by reference in its entirety. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (“>”) symbol in the first column. The word following the “>” symbol is the identifier of the sequence, and the rest of the line is the description (both are optional). There may be no space between the “>” and the first letter of the identifier. It is recommended that all lines of text be shorter than 80 characters. The sequence ends if another line starting with a “>” appears; this indicates the start of another sequence.

[00288] The FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. It is similar to the FASTA format but with quality scores following the sequence data. Both the sequence letter and quality score are encoded with a single ASCII character for brevity. The FASTQ format is a *de facto* standard for storing the output of high throughput sequencing instruments such as the Illumina Genome Analyzer, as described by, for example, Cock *et al.* (“The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants,” *Nucleic Acids Res* 38(6):1767-1771, 2009), which is hereby incorporated by reference in its entirety.

[00289] For FASTA and FASTQ files, meta information includes the description line and not the lines of sequence data. In some embodiments, for FASTQ files, the meta information includes the quality scores. For FASTA and FASTQ files, the sequence data begins after the description line and is present typically using some subset of IUPAC ambiguity codes optionally with “-”. In an embodiment, the sequence data may use the A, T, C, G, and N characters, optionally including “-” or U as-needed (e.g., to represent gaps or uracil).

[00290] In some embodiments, the at least one master sequence read file and the output file are stored as plain text files (e.g., using encoding such as ASCII; ISO/IEC 646; EBCDIC; UTF-8; or UTF-16). A computer system provided by the present disclosure may include a text editor program capable of opening the plain text files. A text editor program may refer to a computer program capable of presenting contents of a text file (such as a plain text file) on a computer screen, allowing a human to edit the text (e.g., using a monitor, keyboard, and mouse). Examples of text editors include, without limitation, Microsoft Word, emacs, pico, vi, BBEdit, and TextWrangler. The text

editor program may be capable of displaying the plain text files on a computer screen, showing the meta information and the sequence reads in a human-readable format (e.g., not binary encoded but instead using alphanumeric characters as they may be used in print or human writing).

[00291] While methods have been discussed with reference to FASTA or FASTQ files, methods and systems of the present disclosure may be used to compress any suitable sequence file format including, for example, files in the Variant Call Format (VCF) format. A typical VCF file may include a header section and a data section. The header contains an arbitrary number of meta-information lines, each starting with characters ‘##’, and a TAB delimited field definition line starting with a single ‘#’ character. The field definition line names eight mandatory columns and the body section contains lines of data populating the columns defined by the field definition line. The VCF format is described by, for example, Danecek *et al.* (“The variant call format and VCF tools,” *Bioinformatics* 27(15):2156-2158, 2011), which is hereby incorporated by reference in its entirety. The header section may be treated as the meta information to write to the compressed files and the data section may be treated as the lines, each of which can be stored in a master file only if unique.

[00292] Some embodiments provide for the assembly of sequence reads. In assembly by alignment, for example, the sequence reads are aligned to each other or aligned to a reference sequence. By aligning each read, in turn to a reference genome, all of the reads are positioned in relationship to each other to create the assembly. In addition, aligning or mapping the sequence read to a reference sequence can also be used to identify variant sequences within the sequence read. Identifying variant sequences can be used in combination with the methods and systems described herein to further aid in the diagnosis or prognosis of a disease or condition, or for guiding treatment decisions.

[00293] In some embodiments, any or all of the steps are automated. Alternatively, methods of the present disclosure may be embodied wholly or partially in one or more dedicated programs, for example, each optionally written in a compiled language such as C++, then compiled and distributed as a binary. Methods of the present disclosure may be implemented wholly or in part as modules within, or by invoking functionality within, existing sequence analysis platforms. In some embodiments, methods of the present disclosure include a number of steps that are all invoked automatically responsive to a single starting queue (e.g., one or a combination of triggering events sourced from human activity, another computer program, or a machine). Thus,

the present disclosure provides methods in which any or the steps or any combination of the steps can occur automatically responsive to a queue. “Automatically” generally means without intervening human input, influence, or interaction (e.g., responsive only to original or pre-queue human activity).

[00294] The methods of the present disclosure may also encompass various forms of output, which includes an accurate and sensitive interpretation of a subject’s nucleic acid sample. The output of retrieval can be provided in the format of a computer file. In some embodiments, the output is a FASTA file, a FASTQ file, or a VCF file. The output may be processed to produce a text file, or an XML file containing sequence data such as a sequence of the nucleic acid aligned to a sequence of the reference genome. In other embodiments, processing yields output containing coordinates or a string describing one or more mutations in the subject nucleic acid relative to the reference genome. Alignment strings may include Simple UnGapped Alignment Report (SUGAR), Verbose Useful Labeled Gapped Alignment Report (VULGAR), and Compact Idiosyncratic Gapped Alignment Report (CIGAR) (as described by, for example, Ning *et al.*, *Genome Research* 11(10):1725-9, 2001, which is hereby incorporated by reference in its entirety). These strings may be implemented, for example, in the Exonerate sequence alignment software from the European Bioinformatics Institute (Hinxton, UK).

[00295] In some embodiments, a sequence alignment is produced—such as, for example, a sequence alignment map (SAM) or binary alignment map (BAM) file—comprising a CIGAR string (the SAM format is described, e.g., by Li *et al.*, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, 25(16):2078-9, 2009, which is hereby incorporated by reference in its entirety). In some embodiments, CIGAR displays or includes gapped alignments one-per-line. CIGAR is a compressed pairwise alignment format reported as a CIGAR string. A CIGAR string may be useful for representing long (e.g., genomic) pairwise alignments. A CIGAR string may be used in SAM format to represent alignments of reads to a reference genome sequence.

[00296] A CIGAR string may follow an established motif. Each character is preceded by a number, giving the base counts of the event. Characters used can include M, I, D, N, and S (M=match; I=insertion; D=deletion; N=gap; S=substitution). The CIGAR string defines the sequence of matches and/or mismatches and deletions (or gaps). For example, the CIGAR string 2MD3M2D2M may indicate that the alignment contains 2 matches, 1 deletion (number 1 is omitted in order to save some space), 3 matches, 2 deletions, and 2 matches.

IV. COMPUTER SYSTEMS

[00297] Methods of the present disclosure can be implemented using, or with the aid of, computer systems. For example, such methods may comprise: partitioning the DNA sample into a plurality of partitions, wherein the plurality of partitions comprises a hypermethylated partition and a hypomethylated partition; tagging the DNA in the hypermethylated and hypomethylated partitions to generate tagged nucleic acids, wherein the tagged nucleic acids comprise molecular barcodes; obtaining sequence reads of molecules from the hypermethylated partition and sequence reads of molecules from the hypomethylated partition, wherein the sequence reads comprise molecular barcode sequence and sample sequence; grouping sequence reads into families based on at least one of (a) the molecular barcode sequences and (b) genomic positions corresponding to the first and last nucleotides of the sample sequence, wherein the families comprise sequence reads derived from a single DNA molecule in the sample; determining a first set of sequences of molecules from the hypermethylated partition and a second set of sequences of molecules from the hypomethylated partition; and calling a plurality of bases based on the first and second sets of sequences; wherein: (i) calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set requires observation of the transition mutation in a greater number of molecules than calling a C to T or G to A transition mutation relative to the reference sequence based on sequences of molecules of the second set; or (ii) C to T or G to A transition mutations are not called relative to a reference sequence based on sequences of molecules of the first set, or C to T or G to A transition mutations are called relative to a reference sequence based on sequences of molecules of the second set without the use of sequences of molecules of the first set, or a C to T or G to A transition mutation is called relative to a reference sequence only if at least one sequence of a molecule of the second set comprises the C to T or G to A transition mutation.

[00298] FIG. 2 shows a computer system 201 that is programmed or otherwise configured to implement the methods of the present disclosure. The computer system 201 can regulate various aspects sample preparation, sequencing, and/or analysis. In some examples, the computer system 201 is configured to perform sample preparation and sample analysis, including nucleic acid sequencing.

[00299] The computer system 201 includes a central processing unit (CPU, also "processor" and "computer processor" herein) 205, which can be a single core or multi core processor, or a plurality

of processors for parallel processing. The computer system 201 also includes memory or memory location 210 (e.g., random-access memory, read-only memory, flash memory), electronic storage unit 215 (e.g., hard disk), communication interface 220 (e.g., network adapter) for communicating with one or more other systems, and peripheral devices 225, such as cache, other memory, data storage, and/or electronic display adapters. The memory 210, storage unit 215, interface 220, and peripheral devices 225 are in communication with the CPU 205 through a communication network or bus (solid lines), such as a motherboard. The storage unit 215 can be a data storage unit (or data repository) for storing data. The computer system 201 can be operatively coupled to a computer network 230 with the aid of the communication interface 220. The computer network 230 can be the Internet, an internet and/or extranet, or an intranet and/or extranet that is in communication with the Internet. The computer network 230 in some cases is a telecommunication and/or data network. The computer network 230 can include one or more computer servers, which can enable distributed computing, such as cloud computing. The computer network 230, in some cases with the aid of the computer system 201, can implement a peer-to-peer network, which may enable devices coupled to the computer system 201 to behave as a client or a server.

[00300] The CPU 205 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 210. Examples of operations performed by the CPU 205 can include fetch, decode, execute, and writeback.

[00301] The storage unit 215 can store files, such as drivers, libraries, and saved programs. The storage unit 215 can store programs generated by users and recorded sessions, as well as output(s) associated with the programs. The storage unit 215 can store user data, e.g., user preferences and user programs. The computer system 201 in some cases can include one or more additional data storage units that are external to the computer system 201, such as located on a remote server that is in communication with the computer system 201 through an intranet or the Internet. Data may be transferred from one location to another using, for example, a communication network or physical data transfer (e.g., using a hard drive, thumb drive, or other data storage mechanism).

[00302] The computer system 201 can communicate with one or more remote computer systems through the network 230. For embodiment, the computer system 201 can communicate with a remote computer system of a user (e.g., operator). Examples of remote computer systems include personal computers (e.g., portable PC), slate or tablet PC's (e.g., Apple® iPad, Samsung® Galaxy

Tab), telephones, Smart phones (e.g., Apple® iPhone, Android-enabled device, Blackberry®), or personal digital assistants. The user can access the computer system 201 via the network 230.

[00303] 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 201, such as, for example, on the memory 210 or electronic storage unit 215. 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 205. In some cases, the code can be retrieved from the storage unit 215 and stored on the memory 210 for ready access by the processor 205. In some situations, the electronic storage unit 215 can be precluded, and machine-executable instructions are stored on memory 210.

[00304] In an aspect, the present disclosure provides a non-transitory computer-readable medium comprising computer-executable instructions which, when executed by at least one electronic processor, perform at least a portion of a method comprising: partitioning the DNA sample into a plurality of partitions, wherein the plurality of partitions comprises a hypermethylated partition and a hypomethylated partition; tagging the DNA in the hypermethylated and hypomethylated partitions to generate tagged nucleic acids, wherein the tagged nucleic acids comprise molecular barcodes; obtaining sequence reads of molecules from the hypermethylated partition and sequence reads of molecules from the hypomethylated partition, wherein the sequence reads comprise molecular barcode sequence and sample sequence; grouping sequence reads into families based on at least one of (a) the molecular barcode sequences and (b) genomic positions corresponding to the first and last nucleotides of the sample sequence, wherein the families comprise sequence reads derived from a single DNA molecule in the sample; determining a first set of sequences of molecules from the hypermethylated partition and a second set of sequences of molecules from the hypomethylated partition; and calling a plurality of bases based on the first and second sets of sequences; wherein: (i) calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set requires observation of the transition mutation in a greater number of molecules than calling a C to T or G to A transition mutation relative to the reference sequence based on sequences of molecules of the second set; or (ii) C to T or G to A transition mutations are not called relative to a reference sequence based on sequences of molecules of the first set, or C to T or G to A transition mutations are called relative to a reference sequence based on sequences of molecules of the second set without the use of sequences

of molecules of the first set, or a C to T or G to A transition mutation is called relative to a reference sequence only if at least one sequence of a molecule of the second set comprises the C to T or G to A transition mutation.

[00305] The code can be pre-compiled and configured for use with a machine have a processor 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.

[00306] Aspects of the systems and methods provided herein, such as the computer system 201, can be embodied in programming. Various aspects of the technology may be thought of as "products" or "articles of manufacture" typically in the form of machine (or processor) executable code and/or associated data that is carried on or embodied in a type of machine readable medium. Machine-executable code can be stored on an electronic storage unit, such memory (e.g., read-only memory, random-access memory, flash memory) or a hard disk. "Storage" type media can include any or all of the tangible memory of the computers, processors or the like, or associated modules thereof, such as various semiconductor memories, tape drives, disk drives and the like, which may provide non-transitory storage at any time for the software programming.

[00307] 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 those 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.

[00308] 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.

[00309] The computer system 201 can include or be in communication with an electronic display that comprises a user interface (UI) for providing, for example, one or more results of sample analysis. Examples of UIs include, without limitation, a graphical user interface (GUI) and web-based user interface.

[00310] Additional details relating to computer systems and networks, databases, and computer program products are also provided in, for example, Peterson, *Computer Networks: A Systems Approach*, Morgan Kaufmann, 5th Ed. (2011), Kurose, *Computer Networking: A Top-Down Approach*, Pearson, 7th Ed. (2016), Elmasri, *Fundamentals of Database Systems*, Addison Wesley, 6th Ed. (2010), Coronel, *Database Systems: Design, Implementation, & Management*, Cengage Learning, 11th Ed. (2014), Tucker, *Programming Languages*, McGraw-Hill Science/Engineering/Math, 2nd Ed. (2006), and Rhoton, *Cloud Computing Architected: Solution Design Handbook*, Recursive Press (2011), each of which is hereby incorporated by reference in its entirety.

V. APPLICATIONS

1. Cancer and Other Diseases

[00311] The present methods can be used to diagnose presence of conditions, particularly cancer, in a subject, to characterize conditions (e.g., staging cancer or determining heterogeneity of a cancer), monitor response to treatment of a condition, effect prognosis risk of developing a condition or subsequent course of a condition. The present disclosure can also be useful in determining the efficacy of a particular treatment option. Successful treatment options may increase the amount of copy number variation or rare mutations detected in subject's blood if the treatment is successful as more cancers may die and shed DNA. In other examples, this may not occur. In another example, perhaps certain treatment options may be correlated with genetic profiles of cancers over time. This correlation may be useful in selecting a therapy.

[00312] Additionally, if a cancer is observed to be in remission after treatment, the present methods can be used to monitor residual disease or recurrence of disease.

[00313] In some embodiments, the methods and systems disclosed herein may be used to identify customized or targeted therapies to treat a given disease or condition in patients based on the classification of a nucleic acid variant as being of somatic or germline origin. Typically, the disease under consideration is a type of cancer. Non-limiting examples of such cancers include biliary tract cancer, bladder cancer, head and neck cancer, transitional cell carcinoma, urothelial carcinoma, brain cancer, gliomas, astrocytomas, breast carcinoma, metaplastic carcinoma, cervical cancer, cervical squamous cell carcinoma, rectal cancer, colorectal carcinoma, colon cancer, hereditary nonpolyposis colorectal cancer, colorectal adenocarcinomas, gastrointestinal stromal tumors (GISTs), endometrial carcinoma, endometrial stromal sarcomas, esophageal cancer, esophageal squamous cell carcinoma, esophageal adenocarcinoma, ocular melanoma, uveal melanoma, gallbladder carcinomas, gallbladder adenocarcinoma, renal cell carcinoma, clear cell renal cell carcinoma, transitional cell carcinoma, urothelial carcinomas, Wilms tumor, leukemia, acute lymphocytic leukemia (ALL), acute myeloid leukemia (AML), chronic lymphocytic leukemia (CLL), chronic myeloid leukemia (CML), chronic myelomonocytic leukemia (CMML), liver cancer, liver carcinoma, hepatoma, hepatocellular carcinoma, cholangiocarcinoma, hepatoblastoma, Lung cancer, non-small cell lung cancer (NSCLC), mesothelioma, B-cell lymphomas, non-Hodgkin lymphoma, diffuse large B-cell lymphoma, Mantle cell lymphoma, T

cell lymphomas, non-Hodgkin lymphoma, precursor T-lymphoblastic lymphoma/leukemia, peripheral T cell lymphomas, multiple myeloma, nasopharyngeal carcinoma (NPC), neuroblastoma, oropharyngeal cancer, oral cavity squamous cell carcinomas, osteosarcoma, ovarian carcinoma, pancreatic cancer, pancreatic ductal adenocarcinoma, pseudopapillary neoplasms, acinar cell carcinomas. Prostate cancer, prostate adenocarcinoma, skin cancer, melanoma, malignant melanoma, cutaneous melanoma, small intestine carcinomas, stomach cancer, gastric carcinoma, gastrointestinal stromal tumor (GIST), uterine cancer, or uterine sarcoma. Type and/or stage of cancer can be detected from genetic variations including mutations, rare mutations, indels, copy number variations, transversions, translocations, inversion, deletions, aneuploidy, partial aneuploidy, polyploidy, chromosomal instability, chromosomal structure alterations, gene fusions, chromosome fusions, gene truncations, gene amplification, gene duplications, chromosomal lesions, DNA lesions, abnormal changes in nucleic acid chemical modifications, abnormal changes in epigenetic patterns, and abnormal changes in nucleic acid 5-methylcytosine.

[00314] Genetic data can also be used for characterizing a specific form of cancer. Cancers are often heterogeneous in both composition and staging. Genetic profile data may allow characterization of specific sub-types of cancer that may be important in the diagnosis or treatment of that specific sub-type. This information may also provide a subject or practitioner clues regarding the prognosis of a specific type of cancer and allow either a subject or practitioner to adapt treatment options in accord with the progress of the disease. Some cancers can progress to become more aggressive and genetically unstable. Other cancers may remain benign, inactive or dormant. The system and methods of this disclosure may be useful in determining disease progression.

[00315] Further, the methods of the disclosure may be used to characterize the heterogeneity of an abnormal condition in a subject. Such methods can include, e.g., generating a genetic profile of extracellular polynucleotides derived from the subject, wherein the genetic profile comprises a plurality of data resulting from copy number variation and rare mutation analyses. In some embodiments, an abnormal condition is cancer. In some embodiments, the abnormal condition may be one resulting in a heterogeneous genomic population. In the example of cancer, some tumors are known to comprise tumor cells in different stages of the cancer. In other examples, heterogeneity may comprise multiple foci of disease. Again, in the example of cancer, there may

be multiple tumor foci, perhaps where one or more foci are the result of metastases that have spread from a primary site.

[00316] The present methods can be used to generate or profile, fingerprint or set of data that is a summation of genetic information derived from different cells in a heterogeneous disease. This set of data may comprise copy number variation, epigenetic variation, and mutation analyses alone or in combination.

[00317] The present methods can be used to diagnose, prognose, monitor or observe cancers, or other diseases. In some embodiments, the methods herein do not involve the diagnosing, prognosing or monitoring a fetus and as such are not directed to non-invasive prenatal testing. In other embodiments, these methodologies may be employed in a pregnant subject to diagnose, prognose, monitor or observe cancers or other diseases in an unborn subject whose DNA and other polynucleotides may co-circulate with maternal molecules.

[00318] Non-limiting examples of other genetic-based diseases, disorders, or conditions that are optionally evaluated using the methods and systems disclosed herein include achondroplasia, alpha-1 antitrypsin deficiency, antiphospholipid syndrome, autism, autosomal dominant polycystic kidney disease, Charcot-Marie-Tooth (CMT), cri du chat, Crohn's disease, cystic fibrosis, Dercum disease, down syndrome, Duane syndrome, Duchenne muscular dystrophy, Factor V Leiden thrombophilia, familial hypercholesterolemia, familial Mediterranean fever, fragile X syndrome, Gaucher disease, hemochromatosis, hemophilia, holoprosencephaly, Huntington's disease, Klinefelter syndrome, Marfan syndrome, myotonic dystrophy, neurofibromatosis, Noonan syndrome, osteogenesis imperfecta, Parkinson's disease, phenylketonuria, Poland anomaly, porphyria, progeria, retinitis pigmentosa, severe combined immunodeficiency (SCID), sickle cell disease, spinal muscular atrophy, Tay-Sachs, thalassemia, trimethylaminuria, Turner syndrome, velocardiofacial syndrome, WAGR syndrome, Wilson disease, or the like.

[00319] In some embodiments, a method described herein comprises detecting a presence or absence of DNA originating or derived from a tumor cell at a preselected timepoint following a previous cancer treatment of a subject previously diagnosed with cancer using a set of sequence information obtained as described herein. The method may further comprise determining a cancer recurrence score that is indicative of the presence or absence of the DNA originating or derived from the tumor cell for the test subject.

[00320] Where a cancer recurrence score is determined, it may further be used to determine a cancer recurrence status. The cancer recurrence status may be at risk for cancer recurrence, e.g., when the cancer recurrence score is above a predetermined threshold. The cancer recurrence status may be at low or lower risk for cancer recurrence, e.g., when the cancer recurrence score is below a predetermined threshold. In particular embodiments, a cancer recurrence score equal to the predetermined threshold may result in a cancer recurrence status of either at risk for cancer recurrence or at low or lower risk for cancer recurrence.

[00321] In some embodiments, a cancer recurrence score is compared with a predetermined cancer recurrence threshold, and the test subject is classified as a candidate for a subsequent cancer treatment when the cancer recurrence score is above the cancer recurrence threshold or not a candidate for therapy when the cancer recurrence score is below the cancer recurrence threshold. In particular embodiments, a cancer recurrence score equal to the cancer recurrence threshold may result in classification as either a candidate for a subsequent cancer treatment or not a candidate for therapy.

[00322] The methods discussed above may further comprise any compatible feature or features set forth elsewhere herein, including in the section regarding methods of determining a risk of cancer recurrence in a test subject and/or classifying a test subject as being a candidate for a subsequent cancer treatment.

2. Methods of determining a risk of cancer recurrence in a test subject and/or classifying a test subject as being a candidate for a subsequent cancer treatment

[00323] In some embodiments, a method provided herein is a method of determining a risk of cancer recurrence in a test subject. In some embodiments, a method provided herein is a method of classifying a test subject as being a candidate for a subsequent cancer treatment.

[00324] Any of such methods may comprise collecting DNA (e.g., originating or derived from a tumor cell) from the test subject diagnosed with the cancer at one or more preselected timepoints following one or more previous cancer treatments to the test subject. The subject may be any of the subjects described herein. The DNA may be cfDNA. The DNA may be obtained from a tissue sample.

[00325] Any of such methods may comprise capturing a plurality of sets of target regions from DNA from the subject, wherein the plurality of target region sets comprises a sequence-variable target region set and an epigenetic target region set, whereby a captured set of DNA molecules is produced. The capturing step may be performed according to any of the embodiments described elsewhere herein.

[00326] In any of such methods, the previous cancer treatment may comprise surgery, administration of a therapeutic composition, and/or chemotherapy.

[00327] Any of such methods may comprise sequencing the captured DNA molecules, whereby a set of sequence information is produced. The captured DNA molecules of the sequence-variable target region set may be sequenced to a greater depth of sequencing than the captured DNA molecules of the epigenetic target region set.

[00328] Any of such methods may comprise detecting a presence or absence of DNA originating or derived from a tumor cell at a preselected timepoint using the set of sequence information. The detection of the presence or absence of DNA originating or derived from a tumor cell may be performed according to any of the embodiments thereof described elsewhere herein.

[00329] Methods of determining a risk of cancer recurrence in a test subject may comprise determining a cancer recurrence score that is indicative of the presence or absence, or amount, of the DNA originating or derived from the tumor cell for the test subject. The cancer recurrence score may further be used to determine a cancer recurrence status. The cancer recurrence status may be at risk for cancer recurrence, e.g., when the cancer recurrence score is above a predetermined threshold. The cancer recurrence status may be at low or lower risk for cancer recurrence, e.g., when the cancer recurrence score is below a predetermined threshold. In particular embodiments, a cancer recurrence score equal to the predetermined threshold may result in a cancer recurrence status of either at risk for cancer recurrence or at low or lower risk for cancer recurrence.

[00330] Methods of classifying a test subject as being a candidate for a subsequent cancer treatment may comprise comparing the cancer recurrence score of the test subject with a predetermined cancer recurrence threshold, thereby classifying the test subject as a candidate for the subsequent cancer treatment when the cancer recurrence score is above the cancer recurrence threshold or not a candidate for therapy when the cancer recurrence score is below the cancer recurrence threshold. In particular embodiments, a cancer recurrence score equal to the cancer recurrence threshold may

result in classification as either a candidate for a subsequent cancer treatment or not a candidate for therapy. In some embodiments, the subsequent cancer treatment comprises chemotherapy or administration of a therapeutic composition.

[00331] Any of such methods may comprise determining a disease-free survival (DFS) period for the test subject based on the cancer recurrence score; for example, the DFS period may be 1 year, 2 years, 3, years, 4 years, 5 years, or 10 years.

[00332] In some embodiments, the set of sequence information comprises sequence-variable target region sequences, and determining the cancer recurrence score may comprise determining at least a first subscore indicative of the amount of SNVs, insertions/deletions, CNVs and/or fusions present in sequence-variable target region sequences.

[00333] In some embodiments, a number of mutations in the sequence-variable target regions chosen from 1, 2, 3, 4, or 5 is sufficient for the first subscore to result in a cancer recurrence score classified as positive for cancer recurrence. In some embodiments, the number of mutations is chosen from 1, 2, or 3.

[00334] In some embodiments, the set of sequence information comprises epigenetic target region sequences, and determining the cancer recurrence score comprises determining a second subscore indicative of the amount of abnormal sequence reads in the epigenetic target region sequences. Abnormal sequence reads may be reads indicative of an epigenetic state different from DNA found in a corresponding sample from a healthy subject (e.g., cfDNA found in a blood sample from a healthy subject, or DNA found in a tissue sample from a healthy subject where the tissue sample is of the same type of tissue as was obtained from the test subject). The abnormal reads may be consistent with epigenetic changes associated with cancer, e.g., methylation of hypermethylation variable target regions and/or perturbed fragmentation of fragmentation variable target regions, where “perturbed” means different from DNA found in a corresponding sample from a healthy subject.

[00335] In some embodiments, a proportion of reads corresponding to the hypermethylation variable target region set and/or fragmentation variable target region set that indicate hypermethylation in the hypermethylation variable target region set and/or abnormal fragmentation in the fragmentation variable target region set greater than or equal to a value in the range of 0.001%-10% is sufficient for the second subscore to be classified as positive for cancer recurrence. The range may be 0.001%-1%, 0.005%-1%, 0.01%-5%, 0.01%-2%, or 0.01%-1%.

[00336] In some embodiments, any of such methods may comprise determining a fraction of tumor DNA from the fraction of reads in the set of sequence information that indicate one or more features indicative of origination from a tumor cell. This may be done for reads corresponding to some or all of the epigenetic target regions, e.g., including one or both of hypermethylation variable target regions and fragmentation variable target regions (hypermethylation of a hypermethylation variable target region and/or abnormal fragmentation of a fragmentation variable target region may be considered indicative of origination from a tumor cell). This may be done for reads corresponding to sequence variable target regions, e.g., reads comprising alterations consistent with cancer, such as SNVs, indels, CNVs, and/or fusions. The fraction of tumor DNA may be determined based on a combination of reads corresponding to epigenetic target regions and reads corresponding to sequence variable target regions.

[00337] Determination of a cancer recurrence score may be based at least in part on the fraction of tumor DNA, wherein a fraction of tumor DNA greater than a threshold in the range of 10^{-11} to 1 or 10^{-10} to 1 is sufficient for the cancer recurrence score to be classified as positive for cancer recurrence. In some embodiments, a fraction of tumor DNA greater than or equal to a threshold in the range of 10^{-10} to 10^{-9} , 10^{-9} to 10^{-8} , 10^{-8} to 10^{-7} , 10^{-7} to 10^{-6} , 10^{-6} to 10^{-5} , 10^{-5} to 10^{-4} , 10^{-4} to 10^{-3} , 10^{-3} to 10^{-2} , or 10^{-2} to 10^{-1} is sufficient for the cancer recurrence score to be classified as positive for cancer recurrence. In some embodiments, the fraction of tumor DNA greater than a threshold of at least 10^{-7} is sufficient for the cancer recurrence score to be classified as positive for cancer recurrence. A determination that a fraction of tumor DNA is greater than a threshold, such as a threshold corresponding to any of the foregoing embodiments, may be made based on a cumulative probability. For example, the sample was considered positive if the cumulative probability that the tumor fraction was greater than a threshold in any of the foregoing ranges exceeds a probability threshold of at least 0.5, 0.75, 0.9, 0.95, 0.98, 0.99, 0.995, or 0.999. In some embodiments, the probability threshold is at least 0.95, such as 0.99.

[00338] In some embodiments, the set of sequence information comprises sequence-variable target region sequences and epigenetic target region sequences, and determining the cancer recurrence score comprises determining a first subscore indicative of the amount of SNVs, insertions/deletions, CNVs and/or fusions present in sequence-variable target region sequences and a second subscore indicative of the amount of abnormal sequence reads in epigenetic target region sequences, and combining the first and second subscores to provide the cancer recurrence

score. Where the first and second subscores are combined, they may be combined by applying a threshold to each subscore independently (e.g., greater than a predetermined number of mutations (e.g., > 1) in sequence-variable target regions, and greater than a predetermined fraction of abnormal (e.g., tumor) reads in epigenetic target regions), or training a machine learning classifier to determine status based on a plurality of positive and negative training samples.

[00339] In some embodiments, a value for the combined score in the range of -4 to 2 or -3 to 1 is sufficient for the cancer recurrence score to be classified as positive for cancer recurrence.

[00340] In any embodiment where a cancer recurrence score is classified as positive for cancer recurrence, the cancer recurrence status of the subject may be at risk for cancer recurrence and/or the subject may be classified as a candidate for a subsequent cancer treatment.

[00341] In some embodiments, the cancer is any one of the types of cancer described elsewhere herein, e.g., colorectal cancer.

3. Therapies and Related Administration

[00342] In certain embodiments, the methods disclosed herein relate to identifying and administering customized therapies to patients given the status of a nucleic acid variant as being of somatic or germline origin. In some embodiments, essentially any cancer therapy (e.g., surgical therapy, radiation therapy, chemotherapy, and/or the like) may be included as part of these methods. Typically, customized therapies include at least one immunotherapy (or an immunotherapeutic agent). Immunotherapy refers generally to methods of enhancing an immune response against a given cancer type. In certain embodiments, immunotherapy refers to methods of enhancing a T cell response against a tumor or cancer.

[00343] In certain embodiments, the status of a nucleic acid variant from a sample from a subject as being of somatic or germline origin may be compared with a database of comparator results from a reference population to identify customized or targeted therapies for that subject. Typically, the reference population includes patients with the same cancer or disease type as the test subject and/or patients who are receiving, or who have received, the same therapy as the test subject. A customized or targeted therapy (or therapies) may be identified when the nucleic variant and the comparator results satisfy certain classification criteria (e.g., are a substantial or an approximate match).

[00344] In certain embodiments, the customized therapies described herein are typically administered parenterally (e.g., intravenously or subcutaneously). Pharmaceutical compositions containing an immunotherapeutic agent are typically administered intravenously. Certain therapeutic agents are administered orally. However, customized therapies (e.g., immunotherapeutic agents, etc.) may also be administered by methods such as, for example, buccal, sublingual, rectal, vaginal, intraurethral, topical, intraocular, intranasal, and/or intraauricular, which administration may include tablets, capsules, granules, aqueous suspensions, gels, sprays, suppositories, salves, ointments, or the like.

[00345] 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 disclosure described herein may be employed in practicing the invention. It is therefore contemplated that the disclosure 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.

[00346] While the foregoing disclosure has been described in some detail by way of illustration and example for purposes of clarity and understanding, it will be clear to one of ordinary skill in the art from a reading of this disclosure that various changes in form and detail can be made without departing from the true scope of the disclosure and may be practiced within the scope of the appended claims. For example, all the methods, systems, computer readable media, and/or component features, steps, elements, or other aspects thereof can be used in various combinations.

[00347] All patents, patent applications, websites, other publications or documents, accession numbers and the like cited herein are incorporated by reference in their entirety for all purposes to the same extent as if each individual item were specifically and individually indicated to be so

incorporated by reference. If different versions of a sequence are associated with an accession number at different times, the version associated with the accession number at the effective filing date of this application is meant. The effective filing date means the earlier of the actual filing date or filing date of a priority application referring to the accession number, if applicable. Likewise, if different versions of a publication, website or the like are published at different times, the version most recently published at the effective filing date of the application is meant, unless otherwise indicated.

VI. EXAMPLES

i) **Observation of increased frequency of artifactual C to T and G to A transition mutations in cfDNA of a hypermethylated partition**

[00348] This example demonstrates that hypermethylated fractions of cell-free DNA contain higher than expected numbers of apparent C to T and G to A transition mutations. Samples from 30 self-declared healthy individuals were collected and each sample was fractionated into at least two partitions comprising a hypermethylated DNA partition and a hypomethylated DNA partition. The partitions (including hypermethylated and hypomethylated DNA partitions) were sequenced and analyzed for error rates wherein the error rate for each molecule satisfies an 80% agreement threshold among read-level bases with at least two sequence reads representing both DNA strands. FIG. 3 shows SNV per-base error rate as a function of specific nucleotide substitution, i.e., A to C, A to G, A to T, C to A, C to G, C to T, G to A, G to C, G to T, T to A, T to C, and T to G. Bar heights represent the mean SNV error rate while the error bars represent the standard error thereof. Light grey bars represent results from the hypermethylated DNA partition, whereas dark grey bars represent results from the hypomethylated DNA partition. The error rate of C to T, and the complementary G to A, substitutions were the most common errors. C to T and G to A substitutions were elevated in the hypermethylated fraction, which indicate a higher rate of chemical damage in hypermethylated molecules.

[00349] Table 6 is a 2 x 2 contingency table showing the number of hypermethylated molecules and hypomethylated molecules with C to T or G to A nucleotide substitutions versus all other substitutions. As described above, substitutions were determined by read-level bases that satisfy an 80% agreement threshold with at least two sequence reads representing both strands. In Table 6, the numbers of determined substitutions are compared to calculated expected substitutions,

which were determined as the row total times the column total divided by the grand total (n). There are significantly more molecules with C to T or G to A substitutions in hypermethylated molecules than hypomethylated molecules than one would expect by chance, as determined by a chi-square test with 1 degree of freedom (p-value of 8.16×10^{-196}).

Table 6

	C>T; G>A		All Other SNVs	
	Observed	Expected	Observed	Expected
Hypermethylated molecules	3611	2604	1335	2342
Hypomethylated molecules	29380	30387	28325	27318

- ii) **Sequencing method in which calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules from a hypermethylated partition requires observation of the transition mutation in a greater number of reads than calling a C to T or G to A transition mutation relative to the reference sequence based on sequences of molecules from a hypomethylated partition**

[00350] This example describes an embodiment of a method to mitigate the impact of artifactual deaminations in hypermethylated partitions on the accuracy of sequence determinations.

[00351] A DNA sample from a subject (e.g., cfDNA, such as human cfDNA) is obtained and at least two partitions comprising hypermethylated and hypomethylated partitions are prepared therefrom as described herein. The partitions (including hypermethylated and hypomethylated partitions) are differentially tagged and then pooled. Target regions of interest (e.g., sequence-variable target regions and epigenetic target regions) are captured using capture probes, then amplified and sequenced, e.g., using a next-generation and/or sequencing-by-synthesis technique.

[00352] Sequence reads are classified as coming from the hypermethylated or hypomethylated partition based on their tag sequences and are grouped according to the original sample molecule from which they originated according to one or more of the tag sequence, the genomic positions corresponding to the first and last nucleotides of the sample sequence, and/or the sequences of a plurality of bases immediately following the 5' tag sequence and immediately preceding the 3' tag sequence. For each group of reads, a sequence of the molecule from which they originated is determined. Sequences of molecules from the hypomethylated partition are mapped to a reference genome sequence and C to T and G to A mutations are identified where the mutation is observed

in at least 2 or 3 sequences of molecules. Sequences from the hypermethylated partition are mapped to a reference genome sequence and C to T and G to A mutations are identified where the mutation is observed in at least 3, 4, or 5 sequences of molecules, where the number of required sequences of molecules is greater than the number of reads required for identifying a C to T or G to A mutation based on sequences of molecules from the hypomethylated partition.

[00353] The results determined in this way have fewer false positive C to T and G to A mutations than a control sequence determination in which the number of required sequences of molecules for identifying a C to T or G to A mutation based on sequences of molecules from the hypomethylated partition is equal to the number of sequences of molecules required for identifying a C to T or G to A mutation based on sequences of molecules from the hypomethylated partition.

iii) Sequencing method in which sequences of molecules from a hypermethylated partition are not used to call C to T or G to A transition mutations

[00354] This example describes another embodiment of a method to mitigate the impact of artifactual deaminations in hypermethylated partitions on the accuracy of sequence determinations.

[00355] A DNA sample from a subject (e.g., cfDNA, such as human cfDNA) is obtained and at least two partitions comprising hypermethylated and hypomethylated partitions are prepared therefrom as described herein. The partitions (including hypermethylated and hypomethylated partitions) are differentially tagged and then pooled. Target regions of interest (e.g., sequence-variable target regions and epigenetic target regions) are captured using capture probes, then amplified and sequenced, e.g., using a next-generation and/or sequencing-by-synthesis technique.

[00356] Sequence reads are classified as coming from the hypermethylated or hypomethylated partition based on their tag sequences and are grouped according to the original sample molecule from which they originated according to one or more of the tag sequence, the genomic coordinates to which the bases immediately following and preceding the 5' and 3' tag sequences correspond, and/or the sequences of a plurality of bases immediately following and preceding the 5' and 3' tag sequences. For each group of reads, a sequence of the molecule from which they originated is determined. Sequences of molecules from the hypomethylated partition are mapped to a reference genome sequence and C to T and G to A mutations are identified where the mutation is observed in at least 2 or 3 sequences of molecules. Sequences from the hypermethylated partition are mapped to a reference genome sequence and are not used to call C to T and G to A mutations relative to the reference genome sequence.

[00357] The results determined in this way have fewer false positive C to T and G to A mutations than a control sequence determination in which the number of required sequences of molecules for identifying a C to T or G to A mutation based on sequences of molecules from the hypomethylated partition is equal to the number of sequences of molecules required for identifying a C to T or G to A mutation based on sequences of molecules from the hypomethylated partition.

iv) Sequencing method in which calling a C to T or G to A transition mutation relative to a reference sequence based on reads from a hypermethylated partition requires observation of the transition mutation in a greater number of reads than calling a C to T or G to A transition mutation relative to the reference sequence based on reads from a hypomethylated partition

[00358] This example describes another embodiment of a method to mitigate the impact of artifactual deaminations in hypermethylated partitions on the accuracy of sequence determinations.

[00359] A DNA sample from a subject (e.g., cfDNA, such as human cfDNA) is obtained and at least two partitions comprising hypermethylated and hypomethylated partitions are prepared therefrom as described herein. The partitions (including hypermethylated and hypomethylated partitions) are differentially tagged and then pooled. Target regions of interest (e.g., sequence-variable target regions and epigenetic target regions) are captured using capture probes, then amplified and sequenced, e.g., using a next-generation and/or sequencing-by-synthesis technique.

[00360] Sequence reads are classified as coming from the hypermethylated or hypomethylated partition based on their tag sequences. Sequences from the hypomethylated partition are mapped to a reference genome sequence and C to T and G to A mutations are identified where the mutation is observed in at least 2 or 3 reads. Sequences from the hypermethylated partition are mapped to a reference genome sequence and C to T and G to A mutations are identified where the mutation is observed in at least 3, 4, or 5 reads, where the number of required reads is greater than the number of reads required for identifying a C to T or G to A mutations based on sequence from the hypomethylated partition.

[00361] The resulting sequence contains fewer false positive C to T and G to A mutations than a control sequence determination in which the number of required reads is for identifying a C to T or G to A mutation based on sequence from the hypomethylated partition is equal to the number of reads required for identifying a C to T or G to A mutation based on sequence from the hypomethylated partition.

v) Sequencing method in which reads from a hypermethylated partition are not

used to call C to T or G to A transition mutations

[00362] This example describes another embodiment of a method to mitigate the impact of artifactual deaminations in hypermethylated partitions on the accuracy of sequence determinations.

[00363] A DNA sample from a subject (e.g., cfDNA, such as human cfDNA) is obtained and at least two partitions comprising hypermethylated and hypomethylated partitions are prepared therefrom as described herein. The partitions (including hypermethylated and hypomethylated partitions) are differentially tagged and then pooled. Target regions of interest (e.g., sequence-variable target regions and epigenetic target regions) are captured using capture probes, then amplified and sequenced, e.g., using a next-generation and/or sequencing-by-synthesis technique.

[00364] Sequence reads are classified as coming from the hypermethylated or hypomethylated partition based on their tag sequences. Sequences from the hypomethylated partition are mapped to a reference genome sequence and apparent C to T and G to A mutations are identified where the mutation is observed in at least 2 or 3 reads. Sequences from the hypermethylated partition are mapped to a reference genome sequence and these reads are not used to call C to T and G to A mutations relative to the reference genome sequence.

[00365] The resulting sequence contains fewer false positive C to T and G to A mutations than a control sequence determination in which the number of required reads is for identifying a C to T or G to A mutation based on sequence from the hypomethylated partition is equal to the number of reads required for identifying a C to T or G to A mutation based on sequence from the hypomethylated partition.

vi) Characterization of target region probe sets with different concentrations of probes for a sequence-variable target region set and probes for an epigenetic target region set

[00366] This example describes an evaluation of the performance of probe sets containing probes for a sequence-variable target region set and probes for an epigenetic target region set as part of an effort to combine epigenetic and genotypic analysis of liquid biopsy cfDNA.

[00367] Samples of cfDNA were processed prior to being contacted with a target region probe set by performing partitioning based on methylation status (thus generating a plurality of partitions comprising hypermethylated and hypomethylated partitions), end repair, ligation with adapters, and amplified by PCR (e.g., using primers targeted to the adapters).

[00368] The processed samples were contacted with target region probe sets comprising probes for a sequence-variable target region set and probes for an epigenetic target region set. The target region probes were in the form of biotinylated oligonucleotides designed to tile the regions of interest. The probes for the sequence-variable target region set had a footprint of about 50 kb and the probes for the epigenetic target region set had a target region footprint of about 500 kb. The probes for the sequence-variable target region set comprised oligonucleotides targeting a selection of regions identified in Tables 3-5 and the probes for the epigenetic target region set comprised oligonucleotides targeting a selection of hypermethylation variable target regions, hypomethylation variable target regions, CTCF binding target regions, transcription start site target regions, focal amplification target regions, and methylation control regions.

[00369] Captured cfDNA isolated in this way was then prepared for sequencing and sequenced using an Illumina HiSeq or NovaSeq sequencer. Results were analyzed with respect to diversity (number of unique families of sequence reads) and read family size (number of individual reads in each family) of the sequence reads corresponding to the probes for the sequence-variable target region set and the probes for the epigenetic target region set. The values reported below were obtained using 70 ng input DNA. 70 ng input is considered a relatively high amount and represents a challenging condition for maintaining desired levels of diversity and family size.

[00370] When the probes for the sequence-variable target region set and the probes for the epigenetic target region set were used in a 1:1 ratio (i.e., the mass per volume concentration of individual oligonucleotides in the two sets was equal), diversity was about 5-10% lower than expected based on input amount for the sequence-variable target regions. This indicates that the sequencing data did not include the expected number of different read families.

[00371] Probe ratios of 2:1 and 5:1 (epigenetic : sequence-variable probe sets) gave greater reductions in diversity relative to the theoretical value for the sequence-variable target regions.

[00372] Probe ratios of 1:2 or 1:5 (epigenetic : sequence-variable probe sets) gave higher levels of diversity for the sequence-variable target regions, which were generally close to the theoretical value, indicating that at these ratios, the presence of the epigenetic target regions were not present in amounts that substantially interfered with generating the expected number of distinct read families from the sequence-variable target regions.

[00373] For the epigenetic target regions, all ratios gave diversity levels substantially lower than the theoretical value. This is not considered problematic, however, given that analysis of

methylation, copy number, and the like for the epigenetic target regions does not require dense and deep sequencing coverage to the same extent as determining the presence or absence of nucleotide substitutions or indels as intended for the sequence-variable regions.

[00374] The sequence reads may be used to determine sequences or sequences of molecules and call mutations essentially as described in any one of examples ii)-v) above to improve accuracy by reducing the frequency of calling false positive C to T and G to A mutations based on reads or sequences of molecules corresponding to the hypermethylated partition.

vii) Detection of cancer using combined epigenetic and sequence-variable target region sets

[00375] Cohorts of cfDNA samples from cancer patients with different stages of cancer from I to IVA (7 stages total) are processed and sequenced as described above in example vi) using probes at the 1:5 (epigenetic : sequence-variable probe sets) ratio. The sequence-variable target region sequences are analyzed by detecting genomic alterations such as SNVs, insertions, deletions and fusions that can be called with enough support to discriminate real tumor variants from technical errors. The epigenetic target region sequences are analyzed independently to detect methylated fragments in regions that have been shown to be differentially methylated in cancer compared to blood cells. Finally, the results of both analyses are combined to produce a final tumor present/absent call to determine whether they showed a profile consistent with cancer at 95% specificity.

[00376] Detection of cancer was 100% sensitive for either approach alone for the stage IIIA and IIIC cohorts. For all but one of the other cohorts, including analysis of the epigenetic target region sequences increased sensitivity by about 10-30%. The one exception was the stage IIB cohort, in which every sample was either a true positive according to both approaches or a false negative according to both approaches.

[00377] Thus, the disclosed methods and compositions can provide captured cfDNA usable for concurrently sequencing epigenetic and sequence-variable target regions to different sequencing depths in sensitive, combined sequence-based and epigenetic detection of cancer.

[00378] The sequence reads may be used to determine sequences essentially as described in any one of examples ii)-v) above to improve accuracy by reducing the frequency of false positive C to T and G to A mutations based on reads or sequences of molecules corresponding to the hypermethylated partition.

CLAIMS

What is claimed is:

1. A method of analyzing a sample of DNA, the method comprising:
 - partitioning the DNA sample into a plurality of partitions, wherein the plurality of partitions comprises a hypermethylated partition and a hypomethylated partition;
 - tagging the DNA in the hypermethylated and hypomethylated partitions to generate tagged nucleic acids, wherein the tagged nucleic acids comprise molecular barcodes;
 - obtaining sequence reads of molecules from the hypermethylated partition and sequence reads of molecules from the hypomethylated partition, wherein the sequence reads comprise molecular barcode sequence and sample sequence;
 - grouping sequence reads into families based on at least one of (a) the molecular barcode sequences and (b) genomic positions corresponding to the first and last nucleotides of the sample sequence, wherein the families comprise sequence reads derived from a single DNA molecule in the sample;
 - determining a first set of sequences of molecules from the hypermethylated partition and a second set of sequences of molecules from the hypomethylated partition; and
 - calling a plurality of bases based on the first and second sets of sequences, wherein:
 - (i) calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set requires observation of the transition mutation in a greater number of molecules than calling a C to T or G to A transition mutation relative to the reference sequence based on sequences of molecules of the second set; or
 - (ii) C to T or G to A transition mutations are not called relative to a reference sequence based on sequences of molecules of the first set, or C to T or G to A transition mutations are called relative to a reference sequence based on sequences of molecules of the second set without the use of sequences of molecules of the first set, or a C to T or G to A transition mutation is called relative to a reference sequence only if at

least one sequence of a molecule of the second set comprises the C to T or G to A transition mutation.

2. The method of the immediately preceding claim, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set requires observation of the transition mutation in a greater number of molecules than calling a C to T or G to A transition mutation relative to the reference sequence based on sequences of molecules of the second set.
3. The method of any one of the preceding claims, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set requires observation of the transition mutation in at least three molecules.
4. The method of the immediately preceding claim, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set requires observation of the transition mutation in at least four molecules.
5. The method of the immediately preceding claim, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set requires observation of the transition mutation in at least five molecules.
6. The method of any one of the preceding claims, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the second set requires observation of the transition mutation in at least two molecules.
7. The method of the immediately preceding claim, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the second set requires observation of the transition mutation in at least three molecules.
8. The method of any one of the preceding claims, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the first set requires observation of the transition mutation in at least two more molecules than does calling a C to T or G to A transition mutation relative to a reference sequence based on sequences of molecules of the second set.

9. The method of any one of the preceding claims, wherein a first threshold is used for calling a C to T or G to A transition based on sequences of molecules of the first set and a second threshold is used for calling a C to T or G to A transition based on sequences of molecules of the second set; the first threshold provides a first level of specificity for calling a C to T or G to A transition; the second threshold provides a second level of specificity for calling a C to T or G to A transition; and the first level of specificity is approximately equal to the second level of specificity, or the first level of specificity is within 10%, 5%, 4%, 3%, 2%, 1%, 0.5%, 0.2%, or 0.1% of the second level of specificity.
10. The method of the immediately preceding claim, wherein the first and second thresholds are specific for C to T and/or G to A transitions.
11. The method of claim 9 or 10, wherein the first and second thresholds are determined from at least one control sample or a plurality of control samples, optionally wherein the at least one control sample or plurality of control samples are from individuals not suspected of having cancer.
12. The method of any one of claims 1-8, wherein a first group of position-specific background error rates are used for a plurality of positions for sequences of molecules of the first set; a second group of position-specific background error rates are used for a plurality of positions for sequences of molecules of the second set; the second group comprises position-specific background error rates higher than the corresponding position-specific background error rates of the first group; and calling a C to T or G to A transition mutation based on sequences of molecules of the first set requires observation of the C to T or G to A transition mutation at a frequency exceeding the corresponding rate from the first group of position-specific background error rates.
13. The method of the immediately preceding claim, wherein calling a C to T or G to A transition mutation based on sequences of molecules of the first set requires observation of the C to T or G to A transition mutation at a frequency exceeding the corresponding rate from the first group of position-specific background error rates by a factor of at least 2, 3, 4, or 5.

14. The method of the immediately preceding claim, wherein calling a C to T or G to A transition mutation based on sequences of molecules of the first set requires observation of the C to T or G to A transition mutation at a frequency exceeding the corresponding rate from the first group of position-specific background error rates by an amount consistent with a confidence level of at least 95%, 98%, 99%, 99.5%, or 99.9%.
15. The method of any one of claims 12-14, wherein the first and second groups of position-specific background error rates are determined from a plurality of control samples, optionally wherein the control samples are from individuals not suspected of having cancer.
16. The method of any one of claims 12-14, wherein the first and second groups of position-specific background error rates were determined using a plurality of control samples, optionally wherein the control samples are from individuals not suspected of having cancer.
17. The method of any one of claims 12-14, wherein the first and second groups of position-specific background error rates were determined using historical data.
18. The method of any one of claims 12-14, wherein the first and second groups of position-specific background error rates were determined using reads and/or sequences of molecules from the hypermethylated and hypomethylated partitions, respectively.
19. The method of any one of the preceding claims, further comprising obtaining sequence reads of molecules from an intermediate partition; determining a third set of sequences of molecules from the intermediate partition; and calling a plurality of bases based on the third set of sequences.
20. The method of the immediately preceding claim, wherein C to T and G to A transition mutations are called based on sequences of the third set less stringently than C to T and G to A transition mutations are called based on sequences of molecules of the first set.
21. The method of the immediately preceding claim, wherein C to T and G to A transition mutations are called based on sequences of the third set in the same way as C to T and G

to A transition mutations are called based on sequences of the second set or more stringently than C to T and G to A transition mutations are called based on sequences of the second set.

22. A method of analyzing a sample of DNA, the method comprising:

obtaining first and second sets of sequence reads from hypermethylated and hypomethylated partitions of the sample, respectively; and

determining a sequence from the first and second sets of sequence reads, wherein:

(i) calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set requires observation of the transition mutation in a greater number of reads than calling a C to T or G to A transition mutation relative to the reference sequence based on reads of the second set; or

(ii) C to T or G to A transition mutations are not called relative to a reference sequence based on reads of the first set, or C to T or G to A transition mutations are called relative to a reference sequence based on sequences of molecules of the second set without the use of sequences of molecules of the first set, or a C to T or G to A transition mutation is called relative to a reference sequence only if at least one sequence of a molecule of the second set comprises the C to T or G to A transition mutation.

23. The method of the immediately preceding claim, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set requires observation of the transition mutation in a greater number of reads than calling a C to T or G to A transition mutation relative to the reference sequence based on reads of the second set.

24. The method of any one of claims 22 or 23, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set requires observation of the transition mutation in at least three reads.

25. The method of the immediately preceding claim, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set requires observation of the transition mutation in at least four reads.

26. The method of the immediately preceding claim, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set requires observation of the transition mutation in at least five reads.
27. The method of any one of claims 22-26, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the second set requires observation of the transition mutation in at least two reads.
28. The method of the immediately preceding claim, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the second set requires observation of the transition mutation in at least three reads.
29. The method of any one of claims 22-28, wherein calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the first set requires observation of the transition mutation in at least two more reads than does calling a C to T or G to A transition mutation relative to a reference sequence based on reads of the second set.
30. The method of any one of the preceding claims, further comprising obtaining a third set of sequence reads from an intermediate partition, wherein the sequence is determined from the third set in addition to the first and second sets.
31. The method of the immediately preceding claim, wherein C to T and G to A transition mutations are called based on reads of the third set less stringently than C to T and G to A transition mutations are called based on reads of the first set.
32. The method of the immediately preceding claim, wherein C to T and G to A transition mutations are called based on reads of the third set in the same way as C to T and G to A transition mutations are called based on reads of the second set.
33. The method of any one of the preceding claims, wherein the DNA of the hypermethylated partition and the DNA of the hypomethylated partition are differentially tagged.

34. The method of any one of the preceding claims, wherein the DNA of the hypermethylated partition and the DNA of the hypomethylated partition are differentially tagged with sequence tags comprising barcodes.
35. The method of any one of the preceding claims, wherein the hypermethylated and hypomethylated partitions were prepared by contacting the DNA of the sample with a methyl binding reagent immobilized on a solid support.
36. The method of the immediately preceding claim, wherein the methyl binding reagent comprises an MBD.
37. The method of claim 36, wherein the methyl binding reagent comprises an MeCP.
38. The method of claim 36, wherein the methyl binding reagent comprises an antibody that binds a methylated nucleotide, optionally wherein the methylated nucleotide is methylated cytosine.
39. The method of any one of claims 35-38, wherein the method comprises contacting the DNA of the sample with the methyl binding reagent immobilized on the solid support and obtaining the hypomethylated partition and hypermethylated partition based on differential binding to the methyl binding reagent.
40. The method of any one of claims 35-39, wherein the method comprises adding differential tags to the DNA of the hypermethylated partition and the DNA of the hypomethylated partition before sequencing.
41. The method of any one of the preceding claims, wherein determining the sequence comprises mapping the first and second sets of sequence reads to a reference sequence to produce mapped sequence reads.
42. The method of any one of the preceding claims, wherein the DNA of the sample or of the hypermethylated and hypomethylated partitions comprises enriched or captured regions of interest.

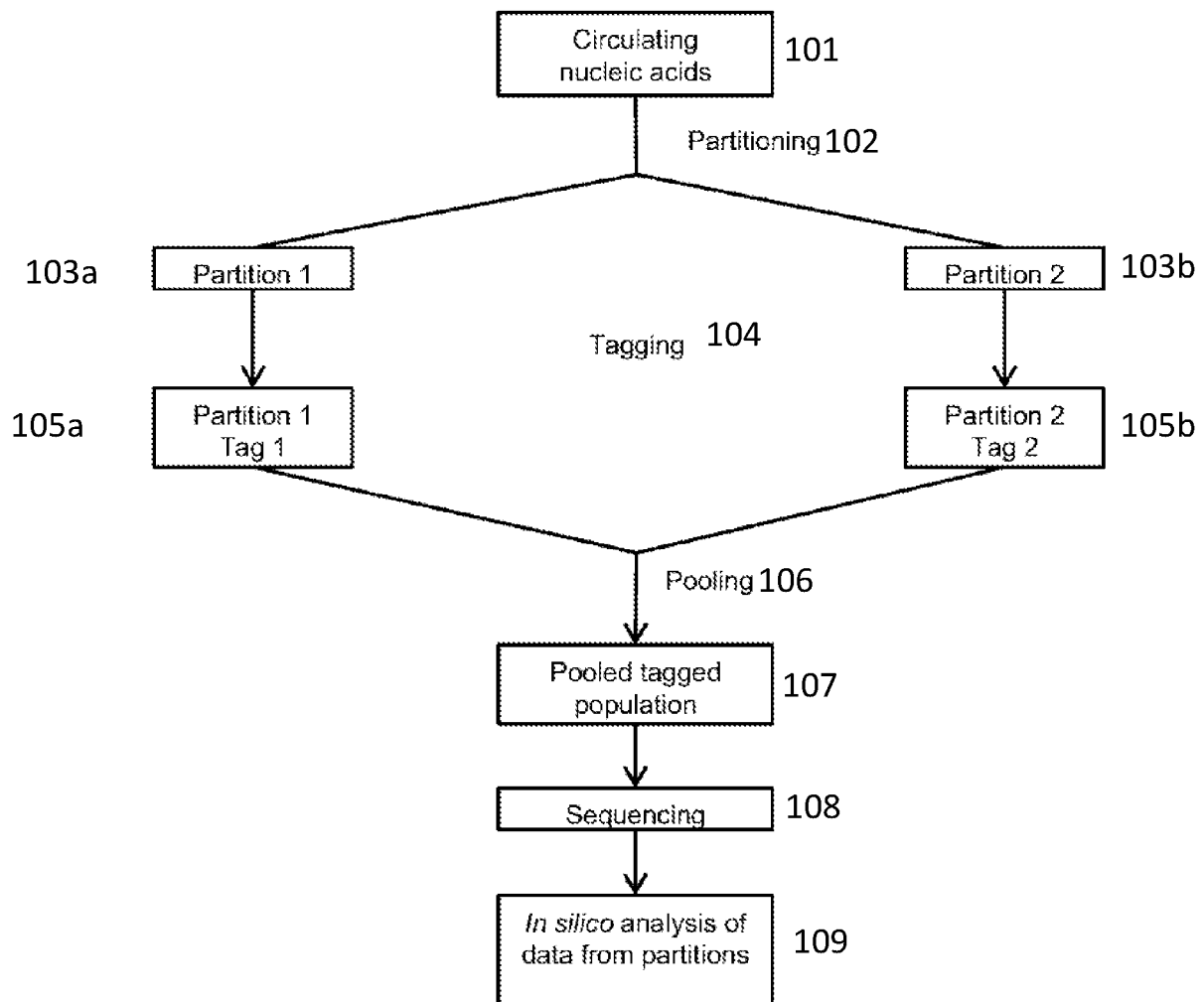
43. The method of any one of the preceding claims, wherein the method comprises enriching the DNA of the sample or of the hypermethylated and hypomethylated partitions for regions of interest or capturing regions of interest from the sample or the hypermethylated and hypomethylated partitions.
44. The method of the immediately preceding claim, wherein enriching or capturing comprises contacting the DNA with a set of target-specific probes, whereby a captured set of DNA molecules is produced.
45. The method of any one of claims 42-44, wherein the regions of interest comprise sequence-variable target regions.
46. The method of the immediately preceding claim, wherein the set of target-specific probes comprises target-binding probes specific for a sequence-variable target set.
47. The method of the immediately preceding claim, wherein the footprint of sequence-variable target region set is at least 25kB or at least 50kB.
48. The method of any one of claims 42-47, wherein the regions of interest comprise epigenetic target regions.
49. The method of the immediately preceding claim, wherein the set of target-specific probes comprises target-binding probes specific for an epigenetic target set.
50. The method of any one of claims 42-49, wherein the regions of interest comprise a sequence-variable target region set and epigenetic target region set.
51. The method of the immediately preceding claim, wherein there are at least 10 regions in the sequence-variable target region set and at least 100 regions in the epigenetic target region set.
52. The method of any one of claims 50-51, wherein the footprint of the epigenetic target region set is at least 2-fold greater than the size of the sequence-variable target region set.

53. The method of the immediately preceding claim, wherein the footprint of the epigenetic target region set is at least 10-fold greater than the size of the sequence-variable target region set.
54. The method of claim 52 or 53, wherein the set of target-specific probes is configured to capture cfDNA corresponding to the sequence-variable target set with a greater capture yield than cfDNA corresponding to the epigenetic target set.
55. The method of any one of claims 50-54, wherein the sequence-variable target region set has a footprint in the range of 10-30 kilobases.
56. The method of any one of claims 50-54, wherein the sequence-variable target region set has a footprint in the range of 30-60 kilobases.
57. The method of any one of claims 50-54, wherein the sequence-variable target region set has a footprint in the range of 60 kilobases to 1 megabase.
58. The method of any one of claims 50-54, wherein the sequence-variable target region set has a footprint in the range of 1-2 megabases.
59. The method of any one of claims 50-58, wherein the epigenetic target region set has a footprint in the range of 0.2-0.8 megabases.
60. The method of any one of claims 50-58, wherein the epigenetic target region set has a footprint in the range of 0.8-1.5 megabases.
61. The method of any one of claims 50-58, wherein the epigenetic target region set has a footprint in the range of 1.5-3 megabases.
62. The method of any one of claims 50-58, wherein the epigenetic target region set has a footprint in the range of 3-8 megabases.
63. The method of any one of claims 50-62, wherein the epigenetic target region set comprises a hypermethylation variable target region set.

64. The method of any one of claims 50-63, wherein the epigenetic target region set comprises a hypomethylation variable target region set.
65. The method of any one of claims 50-64, wherein the epigenetic target region set comprise a fragmentation variable target region set.
66. The method of the immediately preceding claim, wherein the fragmentation variable target region set comprises transcription start site regions.
67. The method of claim 65 or 66, wherein the fragmentation variable target region set comprises CTCF binding regions.
68. The method of any one of claims 50-67, wherein the captured DNA of the sequence-variable target set is sequenced to a greater depth of sequencing than the captured DNA of the epigenetic target region set.
69. The method of the immediately preceding claim, wherein the captured DNA of the sequence-variable target set is sequenced to at least a 2-fold, 3-fold, or 4-fold greater depth of sequencing, or is sequenced to a 4-10-fold or 4-100-fold greater depth of sequencing, than the captured cfDNA molecules of the epigenetic target region set.
70. The method of any one of claims 50-69, wherein the captured DNA of the sequence-variable target set are pooled with the captured DNA of the epigenetic target region set before sequencing.
71. The method of any one of claims 50-70, wherein the captured DNA of the sequence-variable target set and the captured DNA of the epigenetic target region set are sequenced in the same sequencing cell.
72. The method of any one of claims 50-71, wherein the DNA of the hypermethylated and hypomethylated partitions is amplified before capture.
73. The method of any one of the preceding claims, wherein the sample was obtained from a biological tissue or fluid.

74. The method of any one of the preceding claims, wherein the sample was obtained from blood.
75. The method of any one of the preceding claims, wherein the DNA of the sample comprises cell-free DNA.
76. The method of any one of the preceding claims, wherein the DNA of the sample consists essentially of cell-free DNA.
77. The method of any one of the preceding claims, wherein the sample is from a subject having or suspected of having a proliferative disorder or solid tumor.
78. The method of any one of the preceding claims, wherein the sample is from a subject that is undergoing or has undergone treatment for a proliferative disorder or solid tumor.
79. The method of any one of the preceding claims, further comprising determining a likelihood that the subject has a proliferative disorder or solid tumor based on the sequence determined from the sequence reads.
80. The method of any one of the three immediately preceding claims, wherein the proliferative disorder or solid tumor is a cancer.

1/3

**Fig. 1**

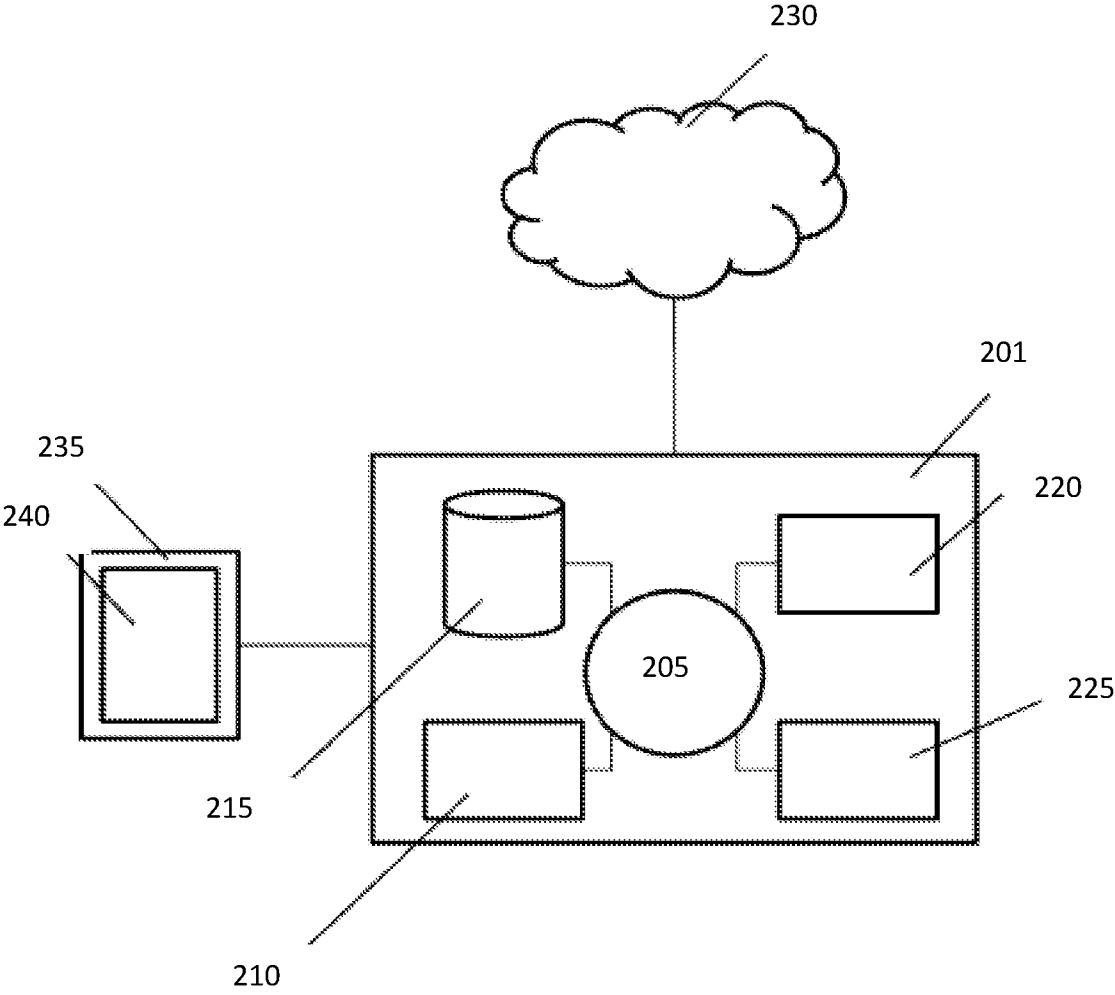


Fig. 2

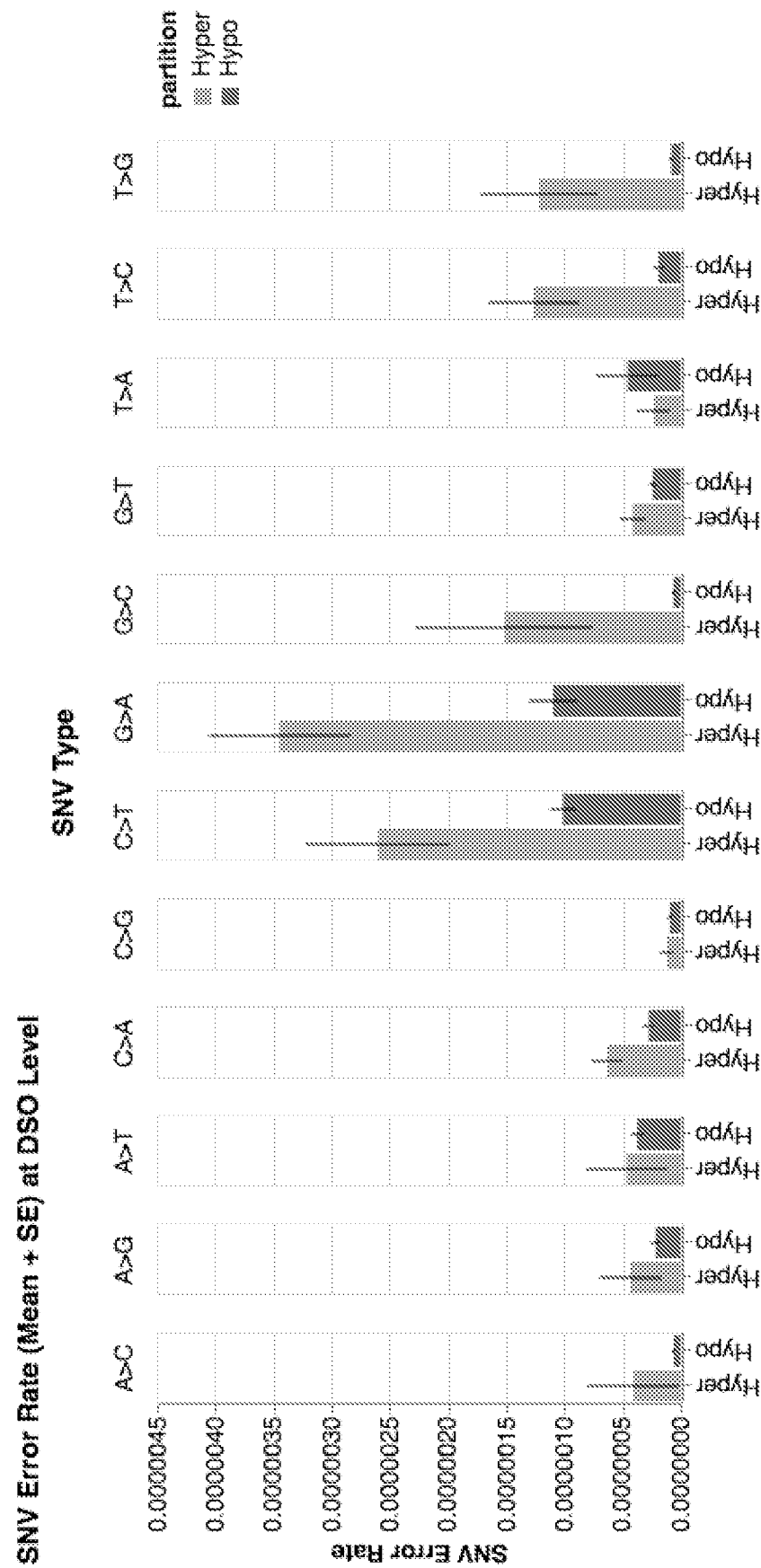


Fig. 3

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2021/030295

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12Q1/6869
ADD.

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)
C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WO 2018/119452 A2 (GUARDANT HEALTH INC [US]) 28 June 2018 (2018-06-28) cited in the application [0023]-[0027],[0050], [0009], [0142] -----	1-80
Y	NEWMAN AARON M ET AL: "Integrated digital error suppression for improved detection of circulating tumor DNA", NATURE BIOTECHNOLOGY, vol. 34, no. 5, 1 May 2016 (2016-05-01), pages 547-555, XP055823049, New York ISSN: 1087-0156, DOI: 10.1038/nbt.3520 abstract, p556-560 "online methods" and Fig. 2a-2d ----- -/-	1-80



Further documents are listed in the continuation of Box C.



See patent family annex.

* 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

14 July 2021

Date of mailing of the international search report

23/07/2021

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Lapopin, Laurence

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2021/030295

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2018/327831 A1 (EATTOCK NICHOLAS [US] ET AL) 15 November 2018 (2018-11-15) claims 1-46 -----	1-80
Y	WO 2017/181146 A1 (GUARDANT HEALTH INC [US]) 19 October 2017 (2017-10-19) [0005]-[0009], [0039] -----	1-80
A	US 2010/144836 A1 (VAN ENGELAND MANON [NL] ET AL) 10 June 2010 (2010-06-10) [0003] -----	1-80

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2021/030295

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2018119452 A2	28-06-2018	AU 2017382439 A1 BR 112019012958 A2 CA 3046007 A1 CN 110325650 A EP 3559270 A2 JP 2020504606 A KR 20190095410 A US 2019390253 A1 WO 2018119452 A2	20-06-2019 26-11-2019 28-06-2018 11-10-2019 30-10-2019 13-02-2020 14-08-2019 26-12-2019 28-06-2018
US 2018327831 A1	15-11-2018	US 2018327831 A1 WO 2018183897 A1	15-11-2018 04-10-2018
WO 2017181146 A1	19-10-2017	EP 3443066 A1 US 2019085406 A1 WO 2017181146 A1	20-02-2019 21-03-2019 19-10-2017
US 2010144836 A1	10-06-2010	CA 2674988 A1 CN 101680025 A EP 2069527 A1 EP 2650377 A1 EP 3434791 A1 JP 5721950 B2 JP 6141896 B2 JP 2010515448 A JP 2015154775 A US 2010144836 A1 US 2015240318 A1 US 2018327859 A1 US 2019032149 A1 WO 2008084219 A1	17-07-2008 24-03-2010 17-06-2009 16-10-2013 30-01-2019 20-05-2015 07-06-2017 13-05-2010 27-08-2015 10-06-2010 27-08-2015 15-11-2018 31-01-2019 17-07-2008