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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 September 2022 (29.09.2022)

WIPO | PCT

(10) International Publication Number
WO 2022/204358 A1

(51) International Patent Classification:
*C12Q 1/68* (2018.01) *G16B 5/20* (2019.01)
*C12Q 1/6876* (2018.01) *G16B 40/00* (2019.01)

(21) International Application Number:
PCT/US2022/021662

(22) International Filing Date:
24 March 2022 (24.03.2022)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/166,641 26 March 2021 (26.03.2021) US

(71) Applicant: **FREENOME HOLDINGS, INC.** [US/US]; 279 East Grand Avenue, 5th Floor, San Francisco, California 94080 (US).

(72) Inventors: **MAHAJAN, Shivani**; 279 East Grand Avenue, 5th Floor, San Francisco, California 94080 (US). **GOULD, Billie**; 279 East Grand Avenue, 5th Floor, San Francisco, California 94080 (US). **ULZ, Peter**; 279 East Grand Avenue, 5th Floor, San Francisco, California 94080 (US).

(74) Agent: **CHOW, Carmen**; WILSON SONSINI GOODRICH & ROSATI, 650 Page Mill Road, Palo Alto, California 94304 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JM, 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,

(54) **Title:** METHODS AND SYSTEMS FOR DETECTING CANCER VIA NUCLEIC ACID METHYLATION ANALYSIS

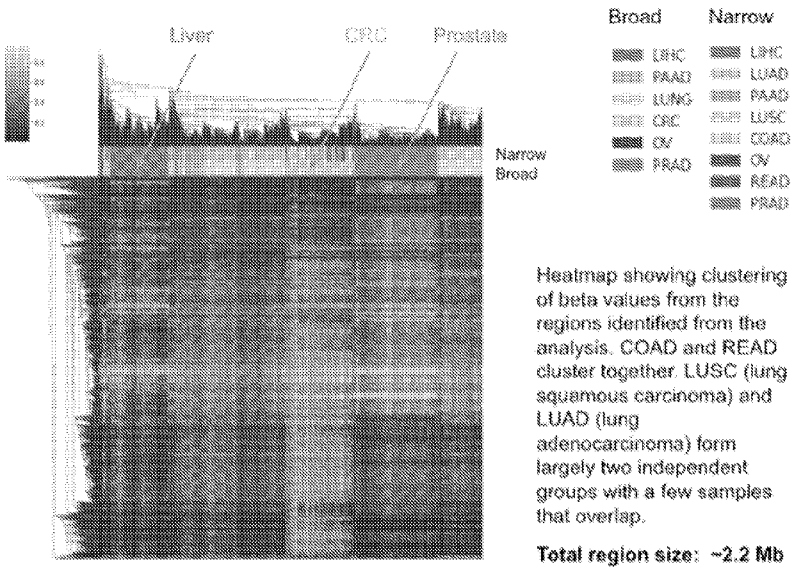



FIG. 2

(57) **Abstract:** The present disclosure provides methods and systems for screening or detecting a tumor or following disease progression that may be applied to cell-free nucleic acids, such as cell-free DNA. The method may use detection of methylation signals within a single sequencing read in identified genomic regions as input features to train a machine learning model and generate a classifier useful for stratifying populations of individuals. The method may comprise extracting DNA from a cell-free sample obtained from a subject, converting the DNA for methylation sequencing, generating sequencing reads, detecting proliferative cell disorder-associated signals in the sequencing information, and training a machine learning model to provide a discriminator capable of distinguishing groups in a subject population such as healthy, cancer, or distinguishing disease subtype or stage. The method may be used for, e.g., predicting, prognosticating, and/or monitoring response to treatment, tumor load, relapse, or cancer development.

WO 2022/204358 A1

*[Continued on next page]*

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).

**Declarations under Rule 4.17:**

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

**Published:**

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

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

# METHODS AND SYSTEMS FOR DETECTING CANCER VIA NUCLEIC ACID METHYLATION ANALYSIS

## CROSS-REFERENCE

**[0001]** This application claims the benefit of U.S. Provisional Patent Application No. 63/166,641, filed on March 26, 2021, the contents of which are incorporated by reference herein.

## INCORPORATION BY REFERENCE

**[0002]** All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference. To the extent publications and patents or patent applications incorporated by reference contradict the disclosure contained in the specification, the specification is intended to supersede and/or take precedence over any such contradictory material.

## BACKGROUND

**[0003]** The present disclosure relates generally to cancer detection and disease monitoring. More particularly, the field relates to cancer-related DNA methylation detection and disease monitoring in early-stage cancer. Cancer screening and monitoring may help to improve outcomes over the past few decades because early detection leads to a better outcome as the cancer may be eliminated before having the opportunity to spread.

**[0004]** A primary issue for any screening tool may be the compromise between false positive and false negative results (or specificity and sensitivity) which lead to unnecessary investigations in the former case, and ineffectiveness in the latter case. An ideal test may be one that has a high Positive Predictive Value (PPV), minimizing unnecessary investigations but detecting the vast majority of cancers. Another key factor is “detection sensitivity”. Distinct from test sensitivity, detection sensitivity is the lower limits of detection with respect to the size of the tumor. Unfortunately, waiting for a tumor to grow large enough to release circulating tumor markers at levels necessary for detection may contradict the goal of treating a tumor at the early stages where treatments are most effective. Hence, there is a need for effective blood-based screens for early-stage cancer based on circulating analytes.

## SUMMARY

**[0005]** The present disclosure provides methods and systems directed to methylation profiling of genes associated with cell proliferative disorder and cancer detection, and disease progression. Further provided are methods and systems directed to methylation profiling of genes associated

with lung, colon, liver, ovarian, pancreatic, prostate, rectal, and breast cell proliferative disorder detection and disease progression.

**[0006]** In an aspect, the present disclosure provides a methylation signature panel characteristic of at least two cell proliferative disorders comprising: six or more methylated genomic regions selected from the group consisting of **Table 1**, wherein the one or more regions are more methylated in a biological sample from an subject having a cell proliferative disorder or cell proliferative disorder subtype and are less methylated in normal tissues and normal blood cells in an subject not having a cell proliferative disorder.

**[0007]** In some embodiments, the biological sample comprises a nucleic acid, DNA, RNA, or cell-free nucleic acid (cfDNA or cfRNA).

**[0008]** In some embodiments, the genomic region is a non-coding region, a coding region, or a non-transcribed or regulator region.

**[0009]** In some embodiments, the signature panel comprises increased methylation in 6 or more, or 12 or more genomic regions in **Table 1**.

**[0010]** In some embodiments, the signature panel comprises increased methylation in six or more methylated genomic regions in **Table 1** that are associated with a type of cancer.

**[0011]** In some embodiments, the biological sample obtained from the subject is selected from the group consisting of body fluids, stool, colonic effluent, urine, blood plasma, blood serum, whole blood, isolated blood cells, cells isolated from the blood, and combinations thereof.

**[0012]** In some embodiments, the cell proliferative disorder is selected from colorectal, prostate, lung, breast, pancreatic, ovarian, uterine, liver, esophagus, stomach, or thyroid cell proliferation.

**[0013]** In some embodiments, the cell proliferative disorder is selected from colon adenocarcinoma, liver hepatocellular carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, ovarian serious cystadenocarcinoma, pancreatic adenocarcinoma, prostate adenocarcinoma, and rectum adenocarcinoma.

**[0014]** In some embodiments, the cell proliferative disorder is selected from stage 1 cancer, stage 2 cancer, stage 3 cancer, or stage 4 cancer.

**[0015]** In some embodiments, the signature panel comprises three or more methylated genomic regions in **Table 1**, four or more methylated genomic regions in **Table 1**, five or more methylated genomic regions in **Table 1,** six or more methylated genomic regions in **Table 1**, seven or more methylated genomic regions in **Table 1**, eight or more methylated genomic regions in **Table 1**, nine or more methylated genomic regions in **Table 1**, ten or more methylated genomic regions in **Table 1**, eleven or more methylated genomic regions in **Table 1**, twelve or more methylated genomic regions in **Table 1**, or thirteen or more methylated genomic regions in **Table 1**.

**[0016]** In an aspect, the present disclosure provides a methylation signature panel characteristic of a tissue of origin for at least two cell proliferative disorders comprising: two or more methylated genomic region signature panels selected from the group consisting of methylated genomic regions in **Tables 2 to 17**, wherein the genomic regions are more methylated in a biological sample from an subject having a cell proliferative disorder or cell proliferative disorder subtype, and are less methylated in normal tissues and normal blood cells in an subject not having a cell proliferative disorder.

**[0017]** In some embodiments, the biological sample is a nucleic acid, DNA, RNA, or cell-free nucleic acid (cfDNA or cfRNA).

**[0018]** In some embodiments, the genomic region is a non-coding region, a coding region, or a non-transcribed or regulator region.

**[0019]** In some embodiments, the signature panel comprises increased methylation in 6 or more, 12 or more genomic regions in **Tables 2 to 17**.

**[0020]** In some embodiments, the signature panel comprises increased methylation in six or more methylated genomic regions in **Tables 2 to 17** that are associated with cancer type and tumor tissue of origin.

**[0021]** In some embodiments, the biological sample obtained from the subject is selected from the group consisting of body fluids, stool, colonic effluent, urine, blood plasma, blood serum, whole blood, isolated blood cells, cells isolated from the blood, and combinations thereof.

**[0022]** In some embodiments, the cell proliferative disorder is selected from colorectal, prostate, lung, breast, pancreatic, ovarian, uterine, liver, esophagus, stomach or thyroid cell proliferation. In some embodiments, the cell proliferative disorder is selected from colon adenocarcinoma, liver hepatocellular carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, ovarian serious cystadenocarcinoma, pancreatic adenocarcinoma, prostate adenocarcinoma, and rectum adenocarcinoma.

**[0023]** In some embodiments, the cell proliferative disorder is selected from stage 1 cancer, stage 2 cancer, stage 3 cancer, or stage 4 cancer.

**[0024]** In some embodiments, the signature panel comprises three or more methylated genomic regions in **Tables 2 to 17**, four or more methylated genomic regions in **Tables 2 to 17**, five or more methylated genomic regions in **Tables 2 to 17,** six or more methylated genomic regions in **Tables 2 to 17**, seven or more methylated genomic regions in **Tables 2 to 17**, eight or more methylated genomic regions in **Tables 2 to 17**, nine or more methylated genomic regions in **Tables 2 to 17**, ten or more methylated genomic regions in **Tables 2 to 17**, eleven or more methylated genomic regions in **Tables 2 to 17**, twelve or more methylated genomic regions in **Tables 2 to 17**, or thirteen or more methylated genomic regions in **Tables 2 to 17**.

**[0025]** In one embodiment, the at least two cell proliferative disorders comprise a combination selected from: colorectal cancer and prostate cancer; colorectal cancer and lung cancer; colorectal cancer and breast cancer; colorectal cancer and liver cancer; colorectal cancer and ovarian cancer; colorectal cancer and pancreatic cancer; prostate cancer and lung cancer; prostate cancer and breast cancer; prostate cancer and liver cancer; prostate cancer and ovarian cancer; prostate cancer and pancreatic cancer; lung cancer and breast cancer; lung cancer and liver cancer; lung cancer and ovarian cancer; lung cancer and pancreatic cancer; breast cancer and liver cancer; breast cancer and ovarian cancer; breast cancer and pancreatic cancer; liver cancer and ovarian cancer; liver cancer and pancreatic cancer; ovarian cancer and pancreatic cancer; colorectal cancer, prostate cancer and lung cancer; colorectal cancer, prostate cancer and breast cancer; colorectal cancer, prostate cancer and liver cancer; colorectal cancer, prostate cancer and ovarian cancer; colorectal cancer, prostate cancer and pancreatic cancer; colorectal cancer, lung cancer and breast cancer; colorectal cancer, lung cancer and liver cancer; colorectal cancer, lung cancer and ovarian cancer; colorectal cancer, lung cancer and pancreatic cancer; colorectal cancer, breast cancer and liver cancer; colorectal cancer, breast cancer and ovarian cancer; colorectal cancer, breast cancer and pancreatic cancer; prostate cancer, liver cancer and ovarian cancer; prostate cancer, liver cancer and pancreatic cancer; prostate cancer, ovarian cancer and pancreatic cancer; and colorectal cancer, prostate cancer, lung cancer, and breast cancer.

**[0026]** In various embodiments, the panel of predetermined methylated genomic regions associated with colorectal cancer tissue of origin is selected from **Tables 2, 3, or 4**.

**[0027]** In various embodiments, the panel of predetermined methylated genomic regions associated with liver cancer tissue of origin is selected from **Tables 5, 6, or 7.**

**[0028]** In various embodiments, the panel of predetermined methylated genomic regions associated with lung cancer tissue of origin is selected from **Tables 8 or 9.**

**[0029]** In various embodiments, the panel of predetermined methylated genomic regions associated with ovarian cancer tissue of origin is selected from **Tables 10, 11, or 12.**

**[0030]** In various embodiments, the panel of predetermined methylated genomic regions associated with pancreatic cancer tissue of origin is selected from **Tables 13 or 14**

**[0031]** In various embodiments, the panel of predetermined methylated genomic regions associated with prostate cancer tissue of origin is selected from **Tables 15, 16, or 17.**

**[0032]** In an aspect, the present disclosure provides a machine learning classifier trained on a panel of predetermined methylated genomic regions associated with 2 or more cancer types wherein the methylated genomic regions are selected from a) **Table 1** and/or b) **Tables 2-17** and combinations thereof.

**[0033]** In another aspect, the present disclosure provides a machine learning classifier capable of distinguishing a population of healthy subjects from subjects with a cell proliferative disorders, comprising:

a) sets of measured values representative of differentially-methylated genomic regions of **Tables 1-17** associated with 2 or more cell proliferative disorders, where the measured values are obtained from methylation sequencing data from healthy subjects and subjects having a cell proliferative disorder,

b) wherein the measured values are used to generate a set of features corresponding to properties of the differentially-methylated genomic regions and where the features are analyzed using a machine learning or statistical model,

c) wherein the model provides a feature vector useful as a classifier capable of distinguishing a population of healthy subjects from subjects having a cell proliferative disorder.

**[0034]** In one embodiment, the sets of measured values describe characteristics of the methylated regions selected from the group consisting of: base wise methylation percent for CpG, CHG, CHH, conversion efficiency (100-mean methylation percent for CHH), hypomethylated blocks, methylation levels (global mean methylation for CPG, CHH, CHG, fragment length, fragment midpoint, and methylation levels in one or more genomic regions such as chrM, LINE1, or ALU), number of methylated CpGs per fragment, fraction of CpG methylation to total CpG per fragment, fraction of CpG methylation to total CpG per region, fraction of CpG methylation to total CpG in panel, dinucleotide coverage (normalized coverage of dinucleotide), evenness of coverage (unique CpG sites at 1x and 10x mean genomic coverage (for S4 runs), mean CpG coverage (depth) globally, and mean coverage at CpG islands (CGI), CGI shelves, CGI shores.

**[0035]** In some embodiments, the panel comprises part of a trained machine learning classifier to classify a subject as having cancer and/or localizing tissue of origin of a tumor in the subject.

**[0036]** In some embodiments, a machine learning model comprising the classifier is loaded into a memory of a computer system, the machine learning model trained using training vectors obtained from training biological samples, a first subset of the training biological samples identified as having a cell proliferative disorder and a second subset of the training biological samples identified as not having a cell proliferative disorder.

**[0037]** In an aspect, the present disclosure provides a machine learning classifier trained on a panel of predetermined methylated genomic regions associated with 2 or more types of cell proliferative disorder, and having pre-selected sensitivity and specificity for the different types of cell proliferative disorder to be detected using the panel.

**[0038]** In various embodiments, the different types of cell proliferative disorders are selected from colorectal cancer, breast cancer, ovarian cancer, prostate cancer, lung cancer, pancreatic cancer, uterine cancer, liver cancer, esophagus cancer, stomach cancer, thyroid cancer, or bladder cancer.

**[0039]** In one embodiment, the machine learning classifier is tailored to provide pre-selected sensitivity and specificity for the different types of cell proliferative disorder to be detected depending on needs of cancer diagnosis and confirmatory diagnosis for two or more cancers selected from colorectal cancer, breast cancer, ovarian cancer, prostate cancer, lung cancer, pancreatic cancer, uterine cancer, liver cancer, esophagus cancer, stomach cancer, thyroid cancer, or bladder cancer, or combinations thereof, wherein the pre-selected sensitivity for a colorectal cancer associated classification panel is at least 70% sensitivity; the pre-selected specificity for a breast cancer associated classification panel is at least 70% specificity; the pre-selected specificity for an ovarian cancer associated classification panel is at least 90% specificity; the pre-selected specificity for a prostate cancer associated classification panel is at least 70% specificity; the pre-selected specificity for a lung cancer associated classification panel is at least 70% specificity; the pre-selected specificity for a pancreatic cancer associated classification panel is at least 90% specificity; the pre-selected specificity for a uterine cancer associated classification panel is at least 90% specificity; the pre-selected sensitivity for a liver cancer associated classification panel is at least 70% sensitivity; the pre-selected sensitivity for an esophagus cancer associated classification panel is at least 70% sensitivity; the pre-selected sensitivity for a stomach cancer associated classification panel is at least 70% sensitivity; the pre-selected specificity for a thyroid cancer associated classification panel is at least 70% specificity; and the pre-selected sensitivity for a bladder cancer associated classification panel is at least 70% sensitivity selected based on which cancer types are detected by the classification model.

**[0040]** In an aspect, the present disclosure provides a method for determining a methylation profile of a cfDNA sample by obtaining, converting, sequencing cfDNA in a sample with a preselected panel of genomic regions associated with the presence of 2 or more cancer types and calculating a methylation profile of cfDNA corresponding to the preselected panel of genomic regions.

**[0041]** In an aspect, the present disclosure provides a method for determining a methylation profile of a cell-free deoxyribonucleic acid (cfDNA) sample from a subject, comprising:

a) providing conditions capable of converting unmethylated cytosines to uracils in nucleic acid molecules of the cfDNA sample to produce a plurality of converted nucleic acids;

b) contacting the plurality of converted nucleic acids with nucleic acid probes complementary to a pre-identified methylation signature panel of at least two differentially methylated regions selected from the group consisting of differentially methylated regions in **Tables 1-17** to enrich for sequences corresponding to the signature panel;

c) determining nucleic acid sequences of the plurality of converted nucleic acid molecules; and

d) aligning the nucleic acid sequences of the plurality of converted nucleic acid molecules to a reference nucleic acid sequence, thereby determining the methylation profile of the subject.

**[0042]** In another aspect, the present disclosure provides a method for determining a methylation profile of a cfDNA sample from a subject comprising:

a) providing conditions capable of converting unmethylated cytosines to uracils in nucleic acid molecules of a cfDNA sample to produce a plurality of converted nucleic acids;

b) amplifying converted nucleic acids with polymerase chain reaction;

c) probing the converted nucleic acids with nucleic acid probes complementary to a pre-identified methylation signature panel of at least two differentially methylated regions selected from **Tables 1-17** to enrich for sequences corresponding to the signature panel;

d) determining the nucleic acid sequence of the converted nucleic acid molecules at a depth of greater than 5000x, and

e) aligning the nucleic acid sequence of the converted nucleic acid molecules to a reference nucleic acid sequence for the pre-identified panel of CpG loci, to determine the methylation profile of the subject.

**[0043]** In some embodiments, a nucleic acid sequencing library is prepared before the amplification.

**[0044]** In some embodiments, the methylation profile is associated with a cell proliferative disorder and provides classification of a subject as having a cell proliferative disorder.

**[0045]** In some embodiments, a nucleic acid adapter comprising a unique molecular identifier is ligated to unconverted nucleic acids in a cfDNA sample before a).

**[0046]** In some embodiments, the nucleic acid molecules are subjected to cytosine-to-uracil conversion conditions using chemical methods, enzymatic methods or a combination thereof.

**[0047]** In some embodiments, the cfDNA in a biological sample is treated with a reagent selected from the group consisting of bisulfite, hydrogen sulfite, disulfite, and combinations thereof.

**[0048]** In some embodiments, the biological sample obtained from the subject is selected from the group consisting of body fluids, stool, colonic effluent, urine, blood plasma, blood serum, whole blood, isolated blood cells, cells isolated from the blood, and combinations thereof.

**[0049]** In some embodiments, the method comprises applying the measured methylation signature panel from the subject against a database of measured methylation signature panels from normal subjects, wherein the database is stored on a computer system; determining that the subject has an increased risk of having a cell proliferative disorder by measuring a change of at least 15% in the methylation status of the methyl signature panel relative to methylation status from normal subjects.

**[0050]** In some embodiments, the cell proliferative disorder is selected from stage 1 cancer, stage 2 cancer, stage 3 cancer, and stage 4 cancer.

**[0051]** In another aspect, the present disclosure provides a method for detecting a cell proliferative disorder in a biological subject comprising:

a) obtaining methylation sequencing information for a preselected panel of genomic regions associated with the presence of 2 or more different cell proliferative disorder tissue types from a nucleic acid sample from the subject,

b) applying the sequence information from the subject to a classification model trained on a preselected panel of genomic regions associated with the presence of 2 or more cell proliferative disorder types, to identify the presence of a cell proliferative disorder, and if a cell proliferative disorder is detected, and

c) applying sequence information from the subject to a classification model trained on a preselected panel of genomic regions associated with associated with the presence of cell proliferative disorders in different tissue types to determine tissue of origin of the cell proliferative disorder in the subject.

**[0052]** In an aspect, the present disclosure provides a method for detecting a cell proliferative disorder in a subject comprising

a) obtaining methylation sequencing information disorders from a nucleic acid sample from the subject for a preselected panel of genomic regions associated with two or more different cell proliferative disorders,

b) calculating a methylation profile of cfDNA in the sample corresponding to the preselected panel of predetermined methylated genomic regions associated with two or more types of cell proliferative disorders, and

c) applying a machine learning classifier trained on a panel of predetermined methylated genomic regions associated with two or more types of cell proliferative disorder, and having pre-selected sensitivity and specificity for the different types of cell proliferative disorder to be detected using the panel.

**[0053]** In various embodiments, the different types of cell proliferative disorders are selected from colorectal cancer, breast cancer, ovarian cancer, prostate cancer, lung cancer, pancreatic

cancer, uterine cancer, liver cancer, esophagus cancer, stomach cancer, thyroid cancer, or bladder cancer,

**[0054]** In one embodiment, the machine learning classifier is tailored to provide pre-selected sensitivity and specificity for the different types of cell proliferative disorder to be detected depending on needs of cancer diagnosis and confirmatory diagnosis for two or more cancers selected from colorectal cancer, breast cancer, ovarian cancer, prostate cancer, lung cancer, pancreatic cancer, uterine cancer, liver cancer, esophagus cancer, stomach cancer, thyroid cancer, or bladder cancer, or combinations thereof.

**[0001]** In one embodiment, the pre-selected sensitivity for a colorectal cancer associated classification panel is at least 70% sensitivity; the pre-selected specificity for a breast cancer associated classification panel is at least 70% specificity; the pre-selected specificity for an ovarian cancer associated classification panel is at least 90% specificity; the pre-selected specificity for a prostate cancer associated classification panel is at least 70% specificity; the pre-selected specificity for a lung cancer associated classification panel is at least 70% specificity; the pre-selected specificity for a pancreatic cancer associated classification panel is at least 90% specificity; the pre-selected specificity for a uterine cancer associated classification panel is at least 90% specificity; the pre-selected sensitivity for a liver cancer associated classification panel is at least 70% sensitivity; the pre-selected sensitivity for an esophagus cancer associated classification panel is at least 70% sensitivity; the pre-selected sensitivity for a stomach cancer associated classification panel is at least 70% sensitivity; the pre-selected specificity for a thyroid cancer associated classification panel is at least 70% specificity; or the pre-selected sensitivity for a bladder cancer associated classification panel is at least 70% sensitivity selected based on which cancer types are detected by the classification model.

**[0055]** In an aspect, the present disclosure provides a method for detecting a presence or an absence of a cell proliferative disorder in a subject, comprising:

a) providing conditions capable of converting unmethylated cytosines to uracils in nucleic acid molecules of a biological sample obtained or derived from the subject to produce a plurality of converted nucleic acids;

b) contacting the plurality of converted nucleic acids with nucleic acid probes complementary to a pre-identified methylation signature panel of at least two differentially methylated regions selected from the group consisting of differentially methylated regions in **Tables 1-17** to enrich for sequences corresponding to the signature panel;

c) determining nucleic acid sequences of the converted nucleic acid molecules;

d) aligning the nucleic acid sequences of the plurality of converted nucleic acid molecules to a reference nucleic acid sequence, thereby determining a methylation profile of the subject; and

e) applying a trained machine learning classifier to the methylation profile, wherein the trained machine learning classifier is trained to be capable of distinguishing between healthy subjects and subjects with a cell proliferative disorder to provide an output value associated with presence of a cell proliferative disorder, thereby detecting the presence or the absence of the cell proliferative disorder in the subject.

**[0056]** In another aspect, the present disclosure provides a method for detecting a cell proliferative disorder in a subject, comprising:

a) providing conditions capable of converting unmethylated cytosines to uracils in nucleic acid molecules of a cfDNA sample to produce a plurality of converted nucleic acids;

b) amplifying converted nucleic acids with polymerase chain reaction;

c) probing the converted nucleic acids with nucleic acid probes complementary to a pre-identified methylation signature panel of at least two differentially methylated regions selected from **Tables 1-17** to enrich for sequences corresponding to the signature panel;

d) determining the nucleic acid sequence of the converted nucleic acid molecules at a depth of greater than 5000x, and

e) aligning the nucleic acid sequence of the converted nucleic acid molecules to a reference nucleic acid sequence for the pre-identified panel of CpG loci, to determine the methylation profile of the subject, and

f) analyzing the methylation profile using a machine learning model trained to be capable of distinguishing between healthy subjects and subjects with a cell proliferative disorder to provide an output value associated with presence of a cell proliferative disorder, thereby indicating the presence of a cell proliferative disorder in the subject.

**[0057]** In some embodiments, the biological sample obtained from the subject is selected from the group consisting of body fluids, stool, colonic effluent, urine, blood plasma, blood serum, whole blood, isolated blood cells, cells isolated from the blood, and combinations thereof.

**[0058]** In some embodiments, the method comprises applying the measured methylation signature panel from the subject against a database of measured methylation signature panels from normal subjects, wherein the database is stored on a computer system; determining that the subject has an increased risk of having a cell proliferative disorder by measuring a change of at least 15% in the methylation status of the methyl signature panel relative to methylation status from normal subjects.

**[0059]** In some embodiments, the cell proliferative disorder is selected from stage 1 cancer, stage 2 cancer, stage 3 cancer, and stage 4 cancer.

**[0060]** In some embodiments, the method detects pancreatic cancer and is performed in combination with detecting the presence or amount of CA19-9 protein in the biological sample.

**[0061]** In some embodiments, the method detects prostate cancer and is performed in combination with detecting the presence or amount of PSA protein in the biological sample.

**[0062]** In an aspect, the present disclosure provides a system comprising a machine learning model classifier for detecting a cell proliferative disorder, comprising:

a) a computer-readable medium comprising a classifier operable to classify subjects as having the cell proliferative disorder or not having the cell proliferative disorder based on a methylation signature panel of **Tables 1-17** or combinations thereof; and

b) one or more processors for executing instructions stored on the computer-readable medium.

**[0063]** In one embodiment, the system comprises the classifier loaded into a memory of a computer system, the machine learning model trained using training vectors obtained from training biological samples, a first subset of the training biological samples identified as having a cell proliferative disorder and a second subset of the training biological samples identified as not having a cell proliferative disorder.

**[0064]** In some embodiments, the classifier is provided in a system for detecting a cell proliferative disorder comprising:

a) a computer-readable medium comprising a classifier operable to classify the subjects based on a methylation signature panel described herein; and

b) one or more processors for executing instructions stored on the computer-readable medium.

**[0065]** In some embodiments, the system comprises a classification circuit that is configured as a machine learning classifier selected from a deep learning classifier, a neural network classifier, a linear discriminant analysis (LDA) classifier, a quadratic discriminant analysis (QDA) classifier, a support vector machine (SVM) classifier, a random forest (RF) classifier, a linear kernel support vector machine classifier, a first or second order polynomial kernel support vector machine classifier, a ridge regression classifier, an elastic net algorithm classifier, a sequential minimal optimization algorithm classifier, a naive Bayes algorithm classifier, and principal component analysis classifier.

**[0066]** In some embodiments, the computer-readable medium is a non-transitory computer-readable medium comprising machine-executable code that, upon execution by one or more computer processors, implements any of the methods above or elsewhere herein.

**[0067]** In some embodiments, the system comprises one or more computer processors and computer memory coupled thereto. The computer memory comprises machine-executable code that, upon execution by the one or more computer processors, implements any of the methods described herein.

**[0068]** In another aspect, the present disclosure provides a method for monitoring minimal residual disease in a subject previously treated for disease comprising: determining a methylation profile as described herein as a baseline methylation state and repeating an analysis to determine the methylation profile at one or more predetermined time points, wherein a change from baseline indicates a change in the minimal residual disease status at baseline in the subject.

**[0069]** In some embodiments, the minimal residual disease is selected from response to treatment, tumor load, residual tumor post-surgery, relapse, secondary screen, primary screen, and cancer progression.

**[0070]** In another aspect, a method is provided for determining response to treatment.

**[0071]** In another aspect, a method is provided for monitoring tumor load.

**[0072]** In another aspect, a method is provided for detecting residual tumor post-surgery.

**[0073]** In another aspect, a method is provided for detecting relapse.

**[0074]** In another aspect, a method is provided for use as a secondary screen.

**[0075]** In another aspect, a method is provided for use as a primary screen.

**[0076]** In another aspect, a method is provided for monitoring cancer progression.

**[0077]** In some embodiments, the dataset is indicative of the presence or susceptibility of the cancer at a sensitivity of at least about 80%. In some embodiments, the dataset is indicative of the presence or susceptibility of the cancer at a sensitivity of at least about 90%. In some embodiments, the dataset is indicative of the presence or susceptibility of the cancer at a sensitivity of at least about 95%. In some embodiments, the dataset is indicative of the presence or susceptibility of the cancer at a positive predictive value (PPV) of at least about 70%. In some embodiments, the dataset is indicative of the presence or susceptibility of the cancer at a positive predictive value (PPV) of at least about 80%. In some embodiments, the dataset is indicative of the presence or susceptibility of the cancer at a positive predictive value (PPV) of at least about 90%. In some embodiments, the dataset is indicative of the presence or susceptibility of the cancer at a positive predictive value (PPV) of at least about 95%. In some embodiments, the dataset is indicative of the presence or susceptibility of the cancer at a positive predictive value (PPV) of at least about 99%. In some embodiments, the dataset is indicative of the presence or susceptibility of the cancer at a negative predictive value (NPV) of at least about 80%. In some embodiments, the dataset is indicative of the presence or susceptibility of the cancer at a negative predictive value (NPV) of at least about 90%. In some embodiments, the dataset is

indicative of the presence or susceptibility of the cancer at a negative predictive value (NPV) of at least about 95%. In some embodiments, the dataset is indicative of the presence or susceptibility of the cancer at a negative predictive value (NPV) of at least about 99%. In some embodiments, the trained algorithm determines the presence or susceptibility of the cancer of the subject with an Area Under Curve (AUC) of at least about 0.90. In some embodiments, the trained algorithm determines the presence or susceptibility of the cancer of the subject with an Area Under Curve (AUC) of at least about 0.95. In some embodiments, the trained algorithm determines the presence or susceptibility of the cancer of the subject with an Area Under Curve (AUC) of at least about 0.99.

**[0078]** In some embodiments, the method further comprises presenting a report a graphical user interface of an electronic device of a user. In some embodiments, the user is the subject, individual, or patient.

**[0079]** In some embodiments, the method further comprises determining a likelihood of the determination of a presence or susceptibility of cancer in the subject, individual, or patient.

**[0080]** In some embodiments, the trained algorithm (e.g., machine learning model or classifier) comprises a supervised machine learning algorithm. In some embodiments, the supervised machine learning algorithm comprises a deep learning algorithm, a support vector machine (SVM), a neural network, or a Random Forest.

**[0081]** In some embodiments, the method further comprises providing said subject with a therapeutic intervention based at least in part on the methylation profile or analysis, such as a therapeutic intervention to treat a patient with cancer (e.g., chemotherapy, radiotherapy, immunotherapy, or surgery).

**[0082]** In some embodiments, the method further comprises monitoring the presence or susceptibility of the cancer, wherein said monitoring comprises assessing the presence or susceptibility of the cancer of said subject at a plurality of time points, wherein the assessing is based at least on the presence or susceptibility of the cancer determined each of the plurality of time points.

**[0083]** In some embodiments, a difference in the assessment of the presence or susceptibility of the cancer of the subject among the plurality of time points is indicative of one or more clinical indications selected from the group consisting of: (i) a diagnosis of the presence or susceptibility of the cancer of the subject; (ii) a prognosis of the presence or susceptibility of the cancer of the subject; and (iii) an efficacy or non-efficacy of a course of treatment for treating the presence or susceptibility of the cancer of the subject.

**[0084]** In some embodiments, the method further comprises stratifying the cancer of the subject by using the trained algorithm to determine a sub-type of the cancer of the subject from among a plurality of distinct subtypes or stages of cancer.

**[0085]** Another aspect of the present disclosure provides a non-transitory computer readable medium comprising machine executable code that, upon execution by one or more computer processors, implements any of the methods above or elsewhere herein.

**[0086]** Another aspect of the present disclosure provides a system comprising one or more computer processors and computer memory coupled thereto. The computer memory comprises machine executable code that, upon execution by the one or more computer processors, implements any of the methods above or elsewhere herein.

**[0087]** Additional aspects and advantages of the present disclosure will become readily apparent to those skilled in this art from the following detailed description, wherein only illustrative embodiments of the present disclosure are shown and described. As will be realized, the present disclosure is capable of other and different embodiments, and its several details are capable of modifications in various obvious respects, all without departing from the disclosure. Accordingly, the drawings and description are to be regarded as illustrative in nature, and not as restrictive.

## BRIEF DESCRIPTION OF THE DRAWINGS

**[0088]** Examples of the present disclosure will now be described, by way of example only, with reference to the attached Figures. The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings (also "Figure" and "FIG." herein), of which:

**[0089] FIG. 1** provides a schematic of a computer system that is programmed or otherwise configured with machine learning models and classifiers to implement methods provided herein.

**[0090] FIG. 2** provides a heatmap of beta values of these 1681 regions that indicates that these regions may contain useful signal for determining tumor of origin as well. Different tumor types cluster into largely distinct groups.

**[0091] FIG. 3** provides a heatmap of the regions included in the multi-cancer panel. The heatmap shows that even with this smaller subset, there is appropriate separation between the different cancer types.

## DETAILED DESCRIPTION

**[0092]** While various embodiments of the invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions may occur to those skilled in the art without departing from the invention. It should be understood that various alternatives to the embodiments of the invention described herein may be employed.

**[0093]** The present disclosure relates generally to cancer detection and disease monitoring. More particularly, the field relates to cancer-related DNA methylation detection and disease monitoring in early-stage cancer. Cancer screening and monitoring may help to improve outcomes because early detection leads to a better outcome as the cancer may be eliminated before having the opportunity to spread. In the case of colorectal cancer, for instance, the use of colonoscopy may play a role in improving early diagnosis. Unfortunately, challenges arise with colonoscopies particularly due to low patient compliance with regular screening.

**[0094]** A primary issue for any screening tool may be the compromise between false positive and false negative results (or specificity and sensitivity), which lead to unnecessary investigations in the former case, and ineffectiveness in the latter case. An ideal test may be one that has a high Positive Predictive Value (PPV), minimizing unnecessary investigations but detecting the vast majority of cancers. Another key factor is “detection sensitivity”. Distinct from test sensitivity, detection sensitivity is the lower limits of detection with respect to the size of the tumor. Unfortunately, waiting for a tumor to grow large enough to release circulating tumor markers at levels necessary for detection may contradict the goal of treating a tumor at the early stages where treatments are most effective. Hence, there is a need for effective blood-based screens for early-stage cancer based on circulating analytes.

**[0095]** Circulating tumor DNA may be a viable “liquid biopsy” for the detection and informative investigation of tumors in a non-invasive manner. The identification of tumor specific mutations in circulating tumor DNA may be applied to diagnosis of colon, breast, and prostate cancers. However, due to the high background of normal (e.g., non-tumor-derived) DNA present in the circulation, these techniques may be limited in sensitivity.

**[0096]** The detection of tumor-specific methylation in the blood may offer distinct advantages over the detection of mutations. A number of single or multiple methylation biomarkers may be assessed in cancers including colorectal, prostate, lung, breast, pancreatic, ovarian, uterine, liver, esophagus, stomach, or thyroid cancer. These biomarkers may suffer from low sensitivities as the biomarkers may be insufficiently prevalent in the tumors. There remains a need for more sensitive and specific screening tools for detecting early-stage or low tumor-burden cancer tumor signals in relapse and primary screening in at risk populations.

**[0097]** The present disclosure provides methods and systems directed to methylation-profiling of genes associated with cancer detection and disease progression.

**[0098]** In an aspect, the present disclosure provides methods that use a panel of methylated regions useful for the analysis of methylation within a region or gene. Other aspects provide novel uses of the region, gene, and the gene product as well as methods, assays, and kits directed to detecting, differentiating, and distinguishing cell proliferative disorders. The method and nucleic acids provided herein may be used for the analysis of cell proliferative disorders such as adenocarcinomas, adenomas, polyps, squamous cell cancers, carcinoid tumors, sarcomas, and lymphomas.

**[0099]** In some embodiments, the method comprises the use of one or more genes of methylated regions as markers for the differentiation, detection, and distinguishing of cell proliferative disorders. In some embodiments, the method comprises analysis of the methylation status of one or more genes selected from methylated regions described herein and their promoter or regulatory elements.

**[0100]** Methods and systems of the present disclosure may comprise analysis of the methylation state of the CpG dinucleotides within one or more of the genomic sequences according to methylated regions described here and sequences complementary thereto.

## I. DEFINITIONS

**[0101]** As used in the specification and claims, the singular form “a”, “an”, and “the” include plural references unless the context clearly dictates otherwise. For example, the term “a nucleic acid” includes a plurality of nucleic acids, including mixtures thereof.

**[0102]** As used herein, the term “subject” generally refers to an entity or a medium that has testable or detectable genetic information. A subject can be a person, individual, or patient. A subject can be a vertebrate, such as, for example, a mammal. Non-limiting examples of mammals include humans, simians, farm animals, sport animals, rodents, and pets. The subject can be a person that has cancer or is suspected of having cancer. The subject may be displaying a symptom(s) indicative of a health or physiological state or condition of the subject, such as a cancer or other disease, disorder, or condition of the subject. As an alternative, the subject can be asymptomatic with respect to such health or physiological state or condition.

**[0103]** As used herein, the term “sample” generally refers to a biological sample obtained from or derived from one or more subjects. Biological samples may be cell-free biological samples or substantially cell-free biological samples, or may be processed or fractionated to produce cell-free biological samples. For example, cell-free biological samples may include cell-free ribonucleic acid (cfRNA), cell-free deoxyribonucleic acid (cfDNA), cell-free fetal DNA

(cffDNA), plasma, serum, urine, saliva, amniotic fluid, and derivatives thereof. Cell-free biological samples may be obtained or derived from subjects using an ethylenediaminetetraacetic acid (EDTA) collection tube, a cell-free RNA collection tube (e.g., Streck®), or a cell-free DNA collection tube (e.g., Streck®). Cell-free biological samples may be derived from whole blood samples by fractionation. Biological samples or derivatives thereof may contain cells. For example, a biological sample may be a blood sample or a derivative thereof (e.g., blood collected by a collection tube or blood drops).

**[0104]** As used herein, the term "nucleic acid" generally refers to a polymeric form of nucleotides of any length, either deoxyribonucleotides (dNTPs) or ribonucleotides (rNTPs), or analogs thereof. Nucleic acids may have any three-dimensional structure, and may perform any function, known or unknown. Non-limiting examples of nucleic acids include deoxyribonucleic (DNA), ribonucleic acid (RNA), coding or non-coding regions of a gene or gene fragment, loci (locus) defined from linkage analysis, exons, introns, messenger RNA (mRNA), transfer RNA, ribosomal RNA, short interfering RNA (siRNA), short-hairpin RNA (shRNA), micro-RNA (miRNA), ribozymes, cDNA, recombinant nucleic acids, branched nucleic acids, plasmids, vectors, isolated DNA of any sequence, isolated RNA of any sequence, nucleic acid probes, and primers. A nucleic acid may comprise one or more modified nucleotides, such as methylated nucleotides and nucleotide analogs. If present, modifications to the nucleotide structure may be made before or after assembly of the nucleic acid. The sequence of nucleotides of a nucleic acid may be interrupted by non-nucleotide components. A nucleic acid may be further modified after polymerization, such as by conjugation or binding with a reporter agent

**[0105]** As used herein, the term "target nucleic acid" generally refers to a nucleic acid molecule in a starting population of nucleic acid molecules having a nucleotide sequence whose presence, amount, and/or sequence, or changes in one or more of these, are desired to be determined. A target nucleic acid may be any type of nucleic acid, including DNA, RNA, and analogs thereof. As used herein, a "target ribonucleic acid (RNA)" generally refers to a target nucleic acid that is RNA. As used herein, a "target deoxyribonucleic acid (DNA)" generally refers to a target nucleic acid that is DNA.

**[0106]** As used herein, the terms "amplifying" and "amplification" generally refer to increasing the size or quantity of a nucleic acid molecule. The nucleic acid molecule may be single-stranded or double-stranded. Amplification may include generating one or more copies or "amplified product" of the nucleic acid molecule. Amplification may be performed, for example, by extension (e.g., primer extension) or ligation. Amplification may include performing a primer extension reaction to generate a strand complementary to a single-stranded nucleic acid molecule, and in some cases generate one or more copies of the strand and/or the single-stranded

nucleic acid molecule. The term "DNA amplification" generally refers to generating one or more copies of a DNA molecule or "amplified DNA product." The term "reverse transcription amplification" generally refers to the generation of deoxyribonucleic acid (DNA) from a ribonucleic acid (RNA) template via the action of a reverse transcriptase

**[0107]** The term "cell-free nucleic acid (cfNA)", as used herein, generally refers to nucleic acids (such as cell-free RNA ("cfRNA") or cell-free DNA ("cfDNA")) in a biological sample that are not contained in a cell. cfDNA may circulate freely in in a bodily fluid, such as in the bloodstream.

**[0108]** The term "cell-free sample", as used herein, generally refers to a biological sample that is substantially devoid of intact cells. This may be derived from a biological sample that is itself substantially devoid of cells or may be derived from a sample from which cells have been removed. Examples of cell-free samples include those derived from blood, such as serum or plasma; urine; or samples derived from other sources, such as semen, sputum, feces, ductal exudate, lymph, or recovered lavage.

**[0109]** The term "circulating tumor DNA", as used herein, generally refers to cfDNA originating from a tumor.

**[0110]** The term "genomic region", as used herein, generally refers to identified regions of nucleic acid that are identified by their location in the chromosome. In some examples, the genomic regions are referred to by a gene name and encompass coding and non-coding regions associated with that physical region of nucleic acid. As used herein, a gene comprises coding regions (exons), non-coding regions (introns), transcriptional control or other regulatory regions, and promoters. In another example, the genomic region may incorporate an intron or exon or an intron/exon boundary within a named gene.

**[0111]** The term "CpG islands" or "CGI", as used herein, generally refers to a contiguous region of genomic DNA that satisfies the criteria of (1) having a frequency of CpG dinucleotides corresponding to an "Observed/Expected Ratio" greater than about 0.6, and (2) having a "GC Content" greater than about 0.5. CpG islands may be between about 0.2 to about 3 kilobases (kb) in length having a high frequency of CpG sites. CpG islands may be found at or near promoters of about 40% of mammalian genes. CpG islands may also be found outside of mammalian genes. In some examples, CpG islands are found in exons, introns, promoters, enhancers, inhibitors, and transcriptional regulatory elements. CpG islands may tend to occur upstream of so-called "housekeeping genes". CpG islands may have a CpG dinucleotide content of at least about 60% of what would be statistically expected. The occurrence of CpG islands at or upstream of the 5' end of genes may reflect a role in the regulation of transcription.

Methylation of CpG sites within the promoters of genes may lead to silencing. Silencing of tumor suppressors by methylation may be, in turn, a hallmark of a number of human cancers.

**[0112]** The term "CpG shores" or "CGI shores", as used herein, generally refers to regions extending short distances from CpG islands in which methylation may also occur. CpG shores may be found in the region about 0 to 2 kb upstream and downstream of a CpG island.

**[0113]** The term "CpG shelves" or "CGI shelves", as used herein, generally refers to regions extending short distances from CpG shores in which methylation may also occur. CpG shelves may generally be found in the region between about 2 kb and 4 kb upstream and downstream of a CpG island (e.g., extending a further 2 kb out from a CpG shore).

**[0114]** The term "cell proliferative disorder", as used herein, generally refers to a disorder or disease that comprises disordered or aberrant proliferation of cells. In some non-limiting examples, the disorder is colorectal cell proliferation, prostate cell proliferation, lung cell proliferation, breast cell proliferation, pancreatic cell proliferation, ovarian cell proliferation, uterine cell proliferation, liver cell proliferation, esophagus cell proliferation, stomach cell proliferation, or thyroid cell proliferation. In some embodiments, the cell proliferative disorder is colon adenocarcinoma, liver hepatocellular carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, ovarian serous cystadenocarcinoma, pancreatic adenocarcinoma, prostate adenocarcinoma, or rectum adenocarcinoma.

**[0115]** The term "normal" or "healthy", as used herein, generally refers to a cell, tissue, plasma, blood, biological sample, or subject not having a cell proliferative disorder.

**[0116]** The term "epigenetic parameters", as used herein, generally refers to cytosine methylations. Further epigenetic parameters may include, for example, the acetylation of histones which may correlate with the DNA methylation.

**[0117]** The term "genetic parameters", as used herein, generally refers to mutations and polymorphisms of genes and sequences further required for gene regulation. Examples of mutations include insertions, deletions, point mutations, inversions, and polymorphisms such as SNPs (single nucleotide polymorphisms).

**[0118]** The term "hemi-methylation" or "hemimethylation", as used herein, generally refers to the methylation state of a palindromic CpG methylation site, where only a single cytosine in one of the two CpG dinucleotide sequences of the palindromic CpG methylation site is methylated (e.g., 5'-CC$^{M}$GG-3' (top strand): 3'-GGCC-5' (bottom strand)).

**[0119]** The term "hypermethylation", as used herein, generally refers to the average methylation state corresponding to an increased presence of 5-mC at one or a plurality of CpG dinucleotides within a DNA sequence of a test DNA sample, relative to the amount of 5-mC found at

corresponding CpG dinucleotides within a normal control DNA sample. In some embodiments, the test DNA sample is from an individual having a cell proliferative disorder.

**[0120]** The term "hypomethylation", as used herein, generally refers to the average methylation state corresponding to a decreased presence of 5-mC at one or a plurality of CpG dinucleotides within a DNA sequence of a test DNA sample, relative to the amount of 5-mC found at corresponding CpG dinucleotides within a normal control DNA sample. In some embodiments, the test DNA sample is from an individual having a cell proliferative disorder.

**[0121]** The term "methylation state" or "methylation status", as used herein, generally refers to the presence or absence of 5-methylcytosine ("5-mC") at one or a plurality of CpG dinucleotides within a DNA sequence. Methylation states at one or more particular palindromic CpG methylation sites (each having two CpG dinucleotide sequences) within a DNA sequence include "unmethylated," "fully-methylated", and "hemi-methylated."

**[0122]** The term "methylated cytosine", as used herein, generally refers to any methylated forms of the nucleic acid base cytosine that contains a methyl or hydroxymethyl functional group at the 5' position. Methylated cytosines may be regulators of gene transcription in genomic DNA. This term may include 5-methylcytosine and 5-hydroxymethylcytosine.

**[0123]** The term "methylation assay" refers to any assay for determining the methylation state of one or more CpG dinucleotide sequences within a sequence of DNA.

**[0124]** The term "minimal residual disease" or "MRD" refers to the small number of cancer cells in the body after cancer treatment. MRD testing may be performed to determine whether the cancer treatment is working and to guide further treatment plans.

**[0125]** The term "MSP" (Methylation-specific PCR), as used herein, generally refers to a methylation assay, such as that described by Herman et al. Proc. Natl. Acad. Sci. USA 93:9821-9826, 1996, and by U.S. Pat. No. 5,786,146, the contents of each of which are incorporated herein by reference herein in its entirety.

**[0126]** The term "methylation converted" or "converted" nucleic acid, as used herein, generally refers to nucleic acid, such as for example DNA, that has undergone a process used to convert the DNA for methylation sequencing. Examples of conversion processes include reagent-based (such as bisulfite) conversion, enzymatic conversion, or combination conversion (such as TAPS conversion) where unmethylated cytosines are converted into uracil prior to PCR amplification or sequencing. The conversion process may be used in methyl sequencing methods to distinguish between methylated and unmethylated cytosine bases.

**[0127]** The term "region methylated in cancer", as used herein, generally refers to a segment of the genome containing methylation sites (CpG dinucleotides), methylation of which is associated with a malignant cellular state. Methylation of a region may be associated with more

than one different type of cancer, or with one type of cancer specifically. Within this, methylation of a region may be associated with more than one cancer subtype, or with one cancer subtype specifically.

**[0128]** The terms cancer "type" and "subtype" generally are used relatively herein, such that one "type" of cancer, such as breast cancer, may be "subtypes" based on, e.g., stage, morphology, histology, gene expression, receptor profile, mutation profile, aggressiveness, prognosis, malignant characteristics, etc. Likewise, "type" and "subtype" may be applied at a finer level, e.g., to differentiate one histological "type" into "subtypes", e.g., defined according to mutation profile or gene expression. Cancer "stage" may also be used to refer to classification of cancer types based on histological and pathological characteristics relating to disease progression.

## II. ASSAYING SAMPLES

**[0129]** The cell-free biological samples may be obtained or derived from a human subject. The cell-free biological samples may be stored in a variety of storage conditions before processing, such as different temperatures (e.g., at room temperature, under refrigeration or freezer conditions, at 25 °C, at 4 °C, at -18 °C, -20 °C, or at -80 °C) or different suspensions (e.g., EDTA collection tubes, cell-free RNA collection tubes, or cell-free DNA collection tubes).

**[0130]** The cell-free biological sample may be obtained from a subject with a cancer, from a subject that is suspected of having a cancer, or from a subject that does not have or is not suspected of having the cancer.

**[0131]** The cell-free biological sample may be taken before and/or after treatment of a subject with the cancer. Cell-free biological samples may be obtained from a subject during a treatment or a treatment regime. Multiple cell-free biological samples may be obtained from a subject to monitor the effects of the treatment over time. The cell-free biological sample may be taken from a subject known or suspected of having a cancer for which a definitive positive or negative diagnosis is not available via clinical tests. The sample may be taken from a subject suspected of having a cancer. The cell-free biological sample may be taken from a subject experiencing unexplained symptoms, such as fatigue, nausea, weight loss, aches and pains, weakness, or bleeding. The cell-free biological sample may be taken from a subject having explained symptoms. The cell-free biological sample may be taken from a subject at risk of developing a cancer due to factors such as familial history, age, hypertension or pre-hypertension, diabetes or pre-diabetes, overweight or obesity, environmental exposure, lifestyle risk factors (e.g., smoking, alcohol consumption, or drug use), or presence of other risk factors.

**[0132]** The cell-free biological sample may contain one or more analytes capable of being assayed, such as cell-free ribonucleic acid (cfRNA) molecules suitable for assaying to generate

transcriptomic data, cell-free deoxyribonucleic acid (cfDNA) molecules suitable for assaying to generate genomic data, or a mixture or combination thereof. One or more such analytes (e.g., cfRNA molecules and/or cfDNA molecules) may be isolated or extracted from one or more cell-free biological samples of a subject for downstream assaying using one or more suitable assays.

**[0133]** After obtaining a cell-free biological sample from the subject, the cell-free biological sample may be processed to generate datasets indicative of a cancer of the subject. For example, a presence, absence, or quantitative assessment of nucleic acid molecules of the cell-free biological sample at a panel of cancer-associated genomic loci (e.g., quantitative measures of RNA transcripts or DNA at the cancer-associated genomic loci). Processing the cell-free biological sample obtained from the subject may comprise (i) subjecting the cell-free biological sample to conditions that are sufficient to isolate, enrich, or extract a plurality of nucleic acid molecules, and (ii) assaying the plurality of nucleic acid molecules to generate the dataset.

**[0134]** In some embodiments, a plurality of nucleic acid molecules is extracted from the cell-free biological sample and subjected to sequencing to generate a plurality of sequencing reads. The nucleic acid molecules may comprise ribonucleic acid (RNA) or deoxyribonucleic acid (DNA). The nucleic acid molecules (e.g., RNA or DNA) may be extracted from the cell-free biological sample by a variety of methods, such as a FastDNA Kit® protocol from MP Biomedicals, a QIAamp® DNA cell-free biological mini kit from Qiagen, or a cell-free biological DNA isolation kit protocol from Norgen Biotek. The extraction method may extract all RNA or DNA molecules from a sample. Alternatively, the extraction method may selectively extract a portion of RNA or DNA molecules from a sample. Extracted RNA molecules from a sample may be converted to DNA molecules by reverse transcription (RT).

**[0135]** The sequencing may be performed by any suitable sequencing methods, such as massively parallel sequencing (MPS), paired-end sequencing, high-throughput sequencing, next-generation sequencing (NGS), shotgun sequencing, single-molecule sequencing, nanopore sequencing, semiconductor sequencing, pyrosequencing, sequencing-by-synthesis (SBS), sequencing-by-ligation, sequencing-by-hybridization, and RNA-Seq (Illumina).

**[0136]** The sequencing may comprise nucleic acid amplification (e.g., of RNA or DNA molecules). In some embodiments, the nucleic acid amplification is polymerase chain reaction (PCR). A suitable number of rounds of PCR (e.g., PCR, qPCR, reverse-transcriptase PCR, digital PCR, etc.) may be performed to sufficiently amplify an initial amount of nucleic acid (e.g., RNA or DNA) to a desired input quantity for subsequent sequencing. In some cases, the PCR may be used for global amplification of target nucleic acids. This may comprise using adapter sequences that may be first ligated to different molecules followed by PCR amplification using universal primers. PCR may be performed using any of a number of commercial kits, e.g.,

provided by Life Technologies, Affymetrix, Promega, Qiagen, etc. In other cases, only certain target nucleic acids within a population of nucleic acids may be amplified. Specific primers, possibly in conjunction with adapter ligation, may be used to selectively amplify certain targets for downstream sequencing. The PCR may comprise targeted amplification of one or more genomic loci, such as genomic loci associated with cancers. The sequencing may comprise use of simultaneous reverse transcription (RT) and polymerase chain reaction (PCR), such as a OneStep RT-PCR kit protocol by Qiagen, NEB, Thermo Fisher Scientific, or Bio-Rad.

**[0137]** RNA or DNA molecules isolated or extracted from a cell-free biological sample may be tagged, e.g., with identifiable tags, to allow for multiplexing of a plurality of samples. Any number of RNA or DNA samples may be multiplexed. For example, a multiplexed reaction may contain RNA or DNA from at least about 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, or more than 100 initial cell-free biological samples. For example, a plurality of cell-free biological samples may be tagged with sample barcodes such that each DNA molecule may be traced back to the sample (and the subject) from which the DNA molecule originated. Such tags may be attached to RNA or DNA molecules by ligation or by PCR amplification with primers.

**[0138]** After subjecting the nucleic acid molecules to sequencing, suitable bioinformatics processes may be performed on the sequence reads to generate the data indicative of the presence, absence, or relative assessment of the cancer. For example, the sequence reads may be aligned to one or more reference genomes (e.g., a genome of one or more species such as a human genome). The aligned sequence reads may be quantified at one or more genomic loci to generate the datasets indicative of the cancer. For example, quantification of sequences corresponding to a plurality of genomic loci associated with cancers may generate the datasets indicative of the cancer.

**[0139]** The cell-free biological sample may be processed without any nucleic acid extraction. For example, the cancer may be identified or monitored in the subject by using probes configured to selectively enrich nucleic acid (e.g., RNA or DNA) molecules corresponding to the plurality of cancer-associated genomic loci. The probes may be nucleic acid primers. The probes may have sequence complementarity with nucleic acid sequences from one or more of the plurality of cancer-associated genomic loci or genomic regions. The plurality of cancer-associated genomic loci or genomic regions may comprise at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least about 25, at least about 30, at least about 35, at least about 40, at least about 45, at least about 50, at least about 55, at least about 60, at least about 65, at least about 70, at least about 75, at least about 80, at

least about 85, at least about 90, at least about 95, at least about 100, or more distinct cancer-associated genomic loci or genomic regions. The plurality of cancer-associated genomic loci or genomic regions may comprise one or more members (e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, about 25, about 30, about 35, about 40, about 45, about 50, about 55, about 60, about 65, about 70, about 75, about 80, or more) selected from the group listed in **Tables 1-11**. The cancer-associated genomic loci or genomic regions may be associated with various stages or sub-types of cancer (e.g., colorectal cancer).

**[0140]** The probes may be nucleic acid molecules (e.g., RNA or DNA) having sequence complementarity with nucleic acid sequences (e.g., RNA or DNA) of the one or more genomic loci (e.g., cancer-associated genomic loci). These nucleic acid molecules may be primers or enrichment sequences. The assaying of the cell-free biological sample using probes that are selective for the one or more genomic loci (e.g., cancer-associated genomic loci) may comprise use of array hybridization (e.g., microarray-based), polymerase chain reaction (PCR), or nucleic acid sequencing (e.g., RNA sequencing or DNA sequencing). In some embodiments, DNA or RNA may be assayed by one or more of: isothermal DNA/RNA amplification methods (e.g., loop-mediated isothermal amplification (LAMP), helicase dependent amplification (HDA), rolling circle amplification (RCA), recombinase polymerase amplification (RPA)), immunoassays, electrochemical assays, surface-enhanced Raman spectroscopy (SERS), quantum dot (QD)-based assays, molecular inversion probes, droplet digital PCR (ddPCR), CRISPR/Cas-based detection (e.g., CRISPR-typing PCR (ctPCR), specific high-sensitivity enzymatic reporter un-locking (SHERLOCK), DNA endonuclease targeted CRISPR trans reporter (DETECTR), and CRISPR-mediated analog multi-event recording apparatus (CAMERA)), and laser transmission spectroscopy (LTS).

**[0141]** The assay readouts may be quantified at one or more genomic loci (e.g., cancer-associated genomic loci) to generate the data indicative of the cancer. For example, quantification of array hybridization or polymerase chain reaction (PCR) corresponding to a plurality of genomic loci (e.g., cancer-associated genomic loci) may generate data indicative of the cancer. Assay readouts may comprise quantitative PCR (qPCR) values, digital PCR (dPCR) values, digital droplet PCR (ddPCR) values, fluorescence values, etc., or normalized values thereof. The assay may be a home use test configured to be performed in a home setting.

**[0142]** In some embodiments, multiple assays may be used to simultaneously process cell-free biological samples of a subject. For example, a first assay may be used to process a first cell-free biological sample obtained or derived from the subject to generate a first dataset indicative of the cancer; and a second assay different from the first assay may be used to process a second cell-free biological sample obtained or derived from the subject to generate a second dataset

indicative of the cancer. Any or all of the first dataset and the second dataset may then be analyzed to assess the cancer of the subject. For example, a single diagnostic index or diagnosis score can be generated based on a combination of the first dataset and the second dataset. As another example, separate diagnostic indexes or diagnosis scores can be generated based on the first dataset and the second dataset.

**[0143]** The cell-free biological samples may be processed using a methylation-specific assay. For example, a methylation-specific assay can be used to identify a quantitative measure (e.g., indicative of a presence, absence, or relative amount) of methylation each of a plurality of cancer-associated genomic loci in a cell-free biological sample of the subject. The methylation-specific assay may be configured to process cell-free biological samples such as a blood sample or a urine sample (or derivatives thereof) of the subject. A quantitative measure (e.g., indicative of a presence, absence, or relative amount) of methylation of cancer-associated genomic loci in the cell-free biological sample may be indicative of one or more cancers. The methylation-specific assay may be used to generate datasets indicative of the quantitative measure (e.g., indicative of a presence, absence, or relative amount) of methylation of each of a plurality of cancer-associated genomic loci in the cell-free biological sample of the subject.

**[0144]** The methylation-specific assay may comprise, for example, one or more of: a methylation-aware sequencing (e.g., using bisulfite treatment), pyrosequencing, methylation-sensitive single-strand conformation analysis (MS-SSCA), high-resolution melting analysis (HRM), methylation-sensitive single-nucleotide primer extension (MS-SnuPE), base-specific cleavage/MALDI-TOF, microarray-based methylation assay, methylation-specific PCR, targeted bisulfite sequencing, oxidative bisulfite sequencing, mass spectroscopy-based bisulfite sequencing, or reduced representation bisulfite sequence (RRBS).

## III. SIGNATURE PANELS

**[0145]** The present disclosure provides methods and systems to analyze biological samples to obtain measurable features from a combination of hypermethylated regions in DNA in the sample that are associated with the development of cell proliferative disorders to identify a signature panel of regions. The features from the signature panel may be processed using a trained algorithm (e.g., a machine learning model) to create a classifier configured to stratify a population of individuals with a cell proliferative disorder. The methods are characterized by using one or more nucleic acids having methylated regions described in the signature panels which are contacted with a reagent or series of reagents capable of distinguishing between methylated and non-methylated CpG dinucleotides within the identified regions prior to sequencing.

**[0146]** The signature panels described herein generally refer to a collection of targeted regions of genomic DNA that are identified in a cell-free nucleic acid sample and display an increased methylation at cytosine bases in samples associated with a cell proliferative disorder. The formation of signature panels may allow for a quick and specific analysis of specific methylated regions associated with cell proliferative disorders. The signature panel(s) as described and employed in the methods herein may be used for the improved diagnosis, prognosis, treatment selection, and monitoring (e.g., treatment monitoring) of cell proliferative disorders such as cancers.

**[0147]** The signature panels and methods may provide significant improvements over current approaches to detect early-stage cell proliferative disorders from body fluid samples such as whole blood, plasma, or serum.

**[0148]** In some embodiments, the regions methylated in cancer comprise CpG islands. In some embodiments, the regions methylated in cancer comprise CpG shores. In some embodiments, the regions methylated in cancer comprise CpG shelves. In some embodiments, the regions methylated in cancer comprise CpG islands and CpG shores. In some embodiments, the regions methylated in cancer comprise CpG islands, CpG shores, and CpG shelves.

**[0149]** In some embodiments, the regions methylated in cancer comprise CpG islands and sequences about 0 to 4 kb upstream and downstream of the CpG islands. The regions methylated in cancer may also comprise CpG islands and sequences about 0 to 3 kb upstream and downstream, about 0 to 2 kb upstream and downstream, about 0 to 1 kb upstream and downstream, about 0 to 500 base pairs (bp) upstream and downstream, about 0 to 400 bp upstream and downstream, about 0 to 300 bp upstream and downstream, about 0 to 200 bp upstream and downstream, or about 0 to 100 bp upstream and downstream of the CpG islands.

**[0150]** A number of design parameters may be considered in the selection of regions hypermethylated in cancer, according to some examples. In certain examples, the methylation region is about 200 bp, about 300 bp, about 400 bp, or about 500 bp in length. Data for this selection process may be obtained from a variety of sources, such as, e.g., The Cancer Genome Atlas (TCGA), derived by the use of, e.g., Illumina Infinium HumanMethylation450 BeadChip for a wide range of cancers, or from other sources based on, e.g., bisulfite whole genome sequencing, or other methodologies. In some embodiments, “methylation value” (which may be derived from TCGA level 3 methylation data, which is in turn derived from the beta-value, which ranges from about -0.5 to 0.5) may be used to select regions. In some embodiments, the amplification is carried out with primer sets designed to amplify at least one methylation site having a methylation value of below about -0.3 in normal issue. A methylation value may be established in a plurality of normal tissue samples, such as about 4. The methylation value may

be at or below about -0.1, about -0.2, about -0.3, about -0.4, about -0.5, about -0.6, about -0.7, about -0.8, about -0.9, or about -1.0.

**[0151]** In some embodiments, the primer sets are designed to amplify at least one methylation site having a difference between the average methylation value in the cancer and the normal tissue of greater than a predefined threshold, such as about 0.3. In some embodiments, the difference may be greater than about 0.1, about 0.2, about 0.3, about 0.4, about 0.5, about 0.6, about 0.7, about 0.8, about 0.9, or about 1.0. Proximity of other methylation sites that meet this requirement may also play a role in selecting regions, in some examples. In some embodiments, the primer sets include primer pairs amplifying at least one methylation site having at least one methylation site within about 200 bp that also has a methylation value of below about -0.3 in normal issue, and a difference between the average methylation value in the cancer and the normal tissue of greater than about 0.3.

**[0152]** In some examples, target regions may be selected if the methylation in a region is greater than methylation in the same region in samples obtained or derived from one or more healthy individuals (e.g., individuals without cancer). Such selection may be performed manually or computationally. In certain examples, a region may be selected if the region has at least about 5%, about 10%, about 15%, about 20%, about 30%, about 40%, about 50%, about 55%, about 60%, about 65%, about 70%, about 75%, about 80%, about 85%, about 90%, about 95%, about 100%, or more than about 100% more methylation than a region in a sample from a healthy individual. In another example, a region may be selected if the number of reads mapped to the region in a disease sample at a predefined threshold methylated CpG count exceeds the same predefined threshold methylated CpG count for the same region in healthy individual samples. The methylated CpG count used as a baseline threshold in healthy samples may change for a given region, but the number of reads mapping to that region that exceeds the baseline threshold of methylated CpG count for that region in a healthy sample may indicate an important region regardless of the fluctuating threshold CpG count.

**[0153]** In some examples, target regions may be selected for amplification based on the number of samples in the validation set having methylation at that site. For example, a region may be selected if the region is more methylated in at least about 5%, about 10%, about 15%, about 20%, about 25%, about 30%, about 35%, about 40%, about 45%, about 50%, about 55%, about 60%, about 65%, about 70%, about 75%, about 80%, about 85%, about 90%, about 95%, about 96%, about 97%, about 98%, or about 99% of samples tested from disease individuals compared to samples from healthy individuals. For example, regions may be selected if the region is methylated in at least about 75% of tumors tested, including within specific subtypes. For some validations, tumor-derived cell lines may be used for the testing.

**[0154]** The present disclosure further provides a method for conducting an assay to ascertain genetic and/or epigenetic parameters of one or more genes selected from the group consisting of the signature panels described herein, and promoter and regulatory elements of the one or more genes. In some embodiments, the assays according to the following method are used to detect methylation within one or more genes selected from the group consisting of signature panels described herein, wherein said methylated nucleic acids are present in a solution further comprising an excess of background DNA, wherein the background DNA is present in between about 100 to 1,000 times, about 100 to 10,000 times, about 100 to 100,000 times, about 1,000 to 10,000 times, about 1,000 to 100,000 times, or about 10,000 to 100,000 times the concentration of the DNA to be detected. In some embodiments, the concentration of DNA to be detected is greater than about 100,000 times the background DNA concentration. In some embodiments, the method comprises contacting a nucleic acid sample obtained from a subject with at least one reagent or a series of reagents (e.g., that distinguishes between methylated and non-methylated CpG dinucleotides within the target nucleic acid).

**[0155]** A tumor or colon cell proliferative disorder, as described herein, may be selected from colorectal, prostate, lung, breast, pancreatic, ovarian, uterine, liver, esophagus, stomach, or thyroid cell proliferation. In some embodiments, the cell proliferative disorder is selected from colon adenocarcinoma, liver hepatocellular carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, ovarian serious cystadenocarcinoma, pancreatic adenocarcinoma, prostate adenocarcinoma, and rectum adenocarcinoma.

**A. Multi Tissue Type Cancer Marker Detection Panel**

**[0156]** A signature panel comprising informative methylated regions may be selected according to the purpose of the intended assay. For targeted methods, primer pairs may be designed based on the set of intended target regions. **Table 1** shows genomic methylation regions indicative of cancer. The methylation regions described herein are annotated to the human reference genome, e.g., from the Genome Reference Consortium Human Build 38 (GRCh38) (The Cancer Genome Atlas (TCGA)). In some embodiments, the set of regions comprises at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, at least ten, at least eleven, at least twelve, at least thirteen, at least fourteen, at least fifteen, at least sixteen, at least seventeen, at least eighteen, at least nineteen, at least twenty, at least twenty five, at least thirty, at least thirty five, at least forty, at least forty five, at least fifty, at least fifty five, or more of the regions listed in **Table 1**. In some embodiments, the set of regions comprise all the regions listed in **Table 1**.

**[0157]** In some embodiments, the set of methyl regions associated with detection of different cancer types is selected from **Table 1**.

**[0158]** In some embodiments, the cancer panel comprises regions selected from at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, at least ten, at least eleven, at least twelve, at least thirteen, at least fourteen, at least fifteen, at least sixteen, at least seventeen, at least eighteen, at least nineteen, at least twenty, at least twenty five, at least thirty, at least thirty five, at least forty, at least forty five, at least fifty, at least fifty five, or more of the regions listed in **Table 1**. In some embodiments, the cancer panel comprises all the regions listed in **Table 1**.

**Table 1**

| *Closest Gene* | *Chromosome Number* | *Region Start* | *Region Stop* |
|---|---|---|---|
| CRIPAK,UVSSA | chr4 | 1410251 | 1411075 |
| MEOX2 | chr7 | 15685300 | 15688105 |
| LRRN2 | chr1 | 204684180 | 204685942 |
| ZNF177 | chr19 | 9362913 | 9363600 |
| DPYSL2 | chr8 | 26578930 | 26579269 |
| CPNE5 | chr6 | 36839707 | 36841032 |
| PHOX2B | chr4 | 41744645 | 41745232 |
| GLB1L3 | chr11 | 134275700 | 134278165 |
| GRIN2A | chr16 | 10181849 | 10183617 |
| SERPINE2 | chr2 | 224038504 | 224040310 |
| LRRC41 | chr1 | 46302590 | 46304060 |
| TMEM101 | chr17 | 44014457 | 44015264 |
| LRRC41 | chr1 | 46301753 | 46301753 |
| TFAP2E | chr1 | 35576535 | 35578237 |
| ZNF154 | chr19 | 57708612 | 57709440 |
| IRF2BP1 | chr19 | 45876550 | 45877420 |
| LYPLAL1-DT | chr1 | 219173767 | 219174230 |
| SCRG1 | chr4 | 173508370 | 173510040 |
| PDXK | chr21 | 43728573 | 43729711 |
| LOC100286906 | chr7 | 155371315 | 155374020 |
| HOPX | chr4 | 56655120 | 56656970 |
| ITGA4 | chr2 | 181456459 | 181458643 |
| PPP1R16B | chr20 | 38804789 | 38807565 |

| | | | |
|---|---|---|---|
| HOXA1 | chr7 | 27095185 | 27097167 |
| TCF24 | chr8 | 66961153 | 66963484 |
| LOC100130992 | chr10 | 22252193 | 22254000 |
| LOC100286906 | chr7 | 155371315 | 155374020 |
| EMILIN2 | chr18 | 2847184 | 2848342 |
| IRF4 | chr6 | 391101 | 392638 |
| OR1F2P | chr16 | 3187792 | 3189855 |
| MOB3B | chr9 | 27528978 | 27529887 |
| IRF2BP1 | chr19 | 45876550 | 45877420 |
| AKR1B1 | chr7 | 134458200 | 134459533 |
| KLK10 | chr19 | 51018748 | 51020133 |
| ZSCAN12 | chr6 | 28399347 | 28400210 |
| RNVU1-8 | chr1 | 147083835 | 147085351 |
| PPP1R16B | chr20 | 38804789 | 38807565 |
| LOC100130298 | chr8 | 60909799 | 60910469 |
| LOC100507557 | chr6 | 145814298 | 145815785 |
| UBXN10 | chr1 | 20185867 | 20186730 |
| ZNF625 | chr19 | 12155907 | 12156871 |
| LOC392232 | chr8 | 72251140 | 72251956 |
| BASP1-AS1 | chr5 | 17217516 | 17219975 |
| LINC01264 | chr10 | 42932936 | 42933788 |
| KIF7 | chr15 | 89654755 | 89655920 |
| TWIST1 | chr7 | 19112343 | 19112657 |
| SNORD12 | chr20 | 49318252 | 49319561 |
| ARHGEF16 | chr1 | 3393377 | 3394370 |
| SLCO3A1 | chr15 | 91852783 | 91854800 |
| TBX18 | chr6 | 84762839 | 84765672 |
| GTDC1 | chr2 | 143936690 | 143937919 |
| SHISA3 | chr4 | 42397040 | 42398745 |
| AKR1B1 | chr7 | 134458200 | 134459533 |
| HOXA1 | chr7 | 27095185 | 27097167 |
| SPAG6 | chr10 | 22334421 | 22337305 |
| IRF4 | chr6 | 391101 | 392638 |
| PRKCB | chr16 | 23835195 | 23837445 |

| | | | |
|---|---|---|---|
| PFKP | chr10 | 3066025 | 3067776 |
| FLT3 | chr13 | 28099335 | 28101240 |

**[0159]** In some embodiments, the method further comprises quantifying the methylation signals, wherein a number in excess of a predetermined threshold is indicative of a cell proliferative disorder such as cancer. In some embodiments, the quantifying and comparing are carried out independently for each of the sites methylated in a cell proliferative disorder. Accordingly, a count of positive tumor signals may be established for each site. In some embodiments, the method further comprises determining a proportion of the sequencing reads containing tumor signals, wherein the proportion in excess of a threshold is indicative of a cell proliferative disorder. In some embodiments, the determining is carried out independently for each of the sites methylated in a cell proliferative disorder.

**[0160]** The term "threshold", as used herein, generally refers to a value that is selected to discriminate, separate, or distinguish between two populations of subjects. In some embodiments, the threshold discriminates methylation status between a disease (e.g., malignant) state, and a non-disease (e.g., healthy) state. In some embodiments, the threshold discriminates between stages of disease (e.g., stage 1, stage 2, stage 3, or stage 4). Thresholds may be set according to the disease in question, and may be based on earlier analysis, e.g., of a training set or determined computationally on a set of inputs having known characteristic (e.g., healthy, disease, or stage of disease). Thresholds may also be set for a gene region according to the predictive value of methylation at a particular site. Thresholds may be different for each methylation site, and data from multiple sites may be combined in the end analysis.

### B. Tissue of Origin Cancer Marker Detection Panel

**[0161]** In some embodiments, of the foregoing methods, the cancer panel comprises methylated genomic regions associated with tissue of origin (TOO) for a type of cancer. The following panels may be incorporated into machine learning classifiers, methods, and systems to determine tissue of origin of tumor-associated methylation signals in a biological sample.

#### i. Colorectal Cancer

**[0162]** **Table 2** shows colorectal tissue of origin TCGA analysis methylation regions. In some embodiments, a cancer panel comprises one or more of the regions listed in **Table 2**. For example, a cancer panel comprises at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 2**. In some embodiments, a set of probes are directed to sequences selected from at least

one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 2**.

**Table 2**

| *Closest Gene* | *Chromosome Number* | *Region Start* | *Region Stop* |
|---|---|---|---|
| GPC6 | chr13 | 93226527 | 93229026 |
| PLCB1 | chr20 | 8131515 | 8134057 |
| PNPLA5 | chr22 | 43890585 | 43892865 |
| NDUFA5P12 | chr8 | 47188540 | 47189584 |
| GRAMD1B | chr11 | 123357785 | 123358764 |
| RNU4-5P | chr10 | 109456465 | 109457325 |
| LRFN3 | chr19 | 35958969 | 35960233 |
| ABHD17AP6 | chr17 | 20852130 | 20853135 |
| GAMT | chr19 | 1400250 | 1401892 |
| WNT2 | chr7 | 117322020 | 117324934 |

**[0163] Table 3** shows colorectal tissue of origin tissue methylation sequencing methylation regions. In some embodiments, a cancer panel comprises one or more of the regions listed in **Table 3**. For example, a cancer panel comprises at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 3**. In some embodiments, a set of probes are directed to sequences selected from at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 3**.

**Table 3**

| *Closest Gene* | *Chromosome Number* | *Region Start* | *Region Stop* |
|---|---|---|---|
| SLC9A3-AS1 | chr5 | 474858 | 475204 |
| DKK2 | chr4 | 107034039 | 107034696 |
| LOX | chr5 | 122076761 | 122077388 |
| C1QL3 | chr10 | 16521666 | 16521939 |
| CSPG5 | chr3 | 47578577 | 47579390 |
| CCT6B | chr17 | 34961183 | 34962104 |
| ISL2 | chr15 | 76336381 | 76336866 |
| ZSWIM2 | chr2 | 186849081 | 186849342 |
| LINC00599 | chr8 | 9904982 | 9906090 |
| LINC01517 | chr10 | 28722124 | 28722695 |

**[0164]** **Table 4** shows colorectal methylation regions overlapping in tissue data and TCGA analysis. In some embodiments, a cancer panel comprises one or more of the regions listed in **Table 4**. For example, a cancer panel comprises at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 4**. In some embodiments, a set of probes are directed to sequences selected from at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 4**. These regions are associated with presence of cancer and are associated with colorectal tissue, and when combined with the regions in **Table 2** and/or **Table 3**, are supportive of colorectal cancer detection.

**Table 4**

| *Closest Gene* | *Chromosome Number* | *Region Start* | *Region Stop* |
|---|---|---|---|
| CRIPAK,UVSSA | chr4 | 1410251 | 1411075 |
| MEOX2 | chr7 | 15685300 | 15688105 |
| LRRN2 | chr1 | 204684180 | 204685942 |
| ZNF177 | chr19 | 9362913 | 9363600 |
| DPYSL2 | chr8 | 26578930 | 26579269 |
| CPNE5 | chr6 | 36839707 | 36841032 |
| PHOX2B | chr4 | 41744645 | 41745232 |
| GLB1L3 | chr11 | 134275700 | 134278165 |
| GRIN2A | chr16 | 10181849 | 10183617 |
| SERPINE2 | chr2 | 224038504 | 224040310 |

**ii. Liver Cancer**

**[0165]** **Table 5** shows liver tissue of origin TCGA analysis methylation regions. In some embodiments, a cancer panel comprises one or more of the regions listed in **Table 5**. For example, a cancer panel comprises at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 5**. In some embodiments, a set of probes are directed to sequences selected from at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 5**.

**Table 5**

| *Closest Gene* | *Chromosome Number* | *Region Start* | *Region Stop* |
|---|---|---|---|

| | | | |
|---|---|---|---|
| PAK1 | chr11 | 77411719 | 77412043 |
| SLC2A1 | chr1 | 42924164 | 42925230 |
| ECE1 | chr1 | 21289565 | 21291415 |
| MREG | chr2 | 216013030 | 216013935 |
| PPM1N | chr19 | 45497996 | 45499985 |
| MYADM | chr19 | 53867498 | 53868212 |
| MEF2C | chr5 | 88883708 | 88884660 |
| ZNF827 | chr4 | 145938434 | 145939260 |
| ZNF510 | chr9 | 96777400 | 96778505 |
| TRAPPC11 | chr4 | 183722761 | 183723261 |

**[0166]** **Table 6** shows liver tissue of origin tissue methylation sequencing methylation regions. In some embodiments, a cancer panel comprises one or more of the regions listed in **Table 6**. For example, a cancer panel comprises at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 6**. In some embodiments, a set of probes are directed to sequences selected from at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 6**.

**Table 6**

| ***Closest Gene*** | ***Chromosome Number*** | ***Region Start*** | ***Region Stop*** |
|---|---|---|---|
| SLC25A36 | chr3 | 140942215 | 140942681 |
| NXPE3 | chr3 | 101778817 | 101779493 |
| SMG1P3 | chr16 | 21519806 | 21520585 |
| KANK1 | chr9 | 503975 | 504835 |
| NANOS1 | chr10 | 119029236 | 119030177 |
| RBM4 | chr11 | 66638405 | 66639220 |
| EFNB2 | chr13 | 106535692 | 106536388 |
| GPR180 | chr13 | 94601865 | 94602451 |
| SPINT2 | chr19 | 38264214 | 38264616 |
| SLC2A1 | chr1 | 42924235 | 42924973 |

**[0167]** **Table 7** shows liver methylation regions overlapping in tissue data and TCGA analysis. In some embodiments, a cancer panel comprises one or more of the regions listed in **Table 7**.

For example, a cancer panel comprises at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 7**. In some embodiments, a set of probes are directed to sequences selected from at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 7**. These regions are associated with presence of cancer and are associated with liver tissue, and when combined with the regions in **Table 5** and/or **Table 6**, are supportive of liver cancer detection.

**Table 7**

| *Closest Gene* | *Chromosome Number* | *Region Start* | *Region Stop* |
|---|---|---|---|
| GTDC1 | chr2 | 143936690 | 143937919 |
| TCF24 | chr8 | 66961153 | 66963484 |
| SHISA3 | chr4 | 42397040 | 42398745 |
| AKR1B1 | chr7 | 134458200 | 134459533 |
| HOXA1 | chr7 | 27095185 | 27097167 |
| SPAG6 | chr10 | 22334421 | 22337305 |
| IRF4 | chr6 | 391101 | 392638 |
| PRKCB | chr16 | 23835195 | 23837445 |
| PFKP | chr10 | 3066025 | 3067776 |
| FLT3 | chr13 | 28099335 | 28101240 |

**iii. Lung Cancer**

**[0168]** **Table 8** shows lung tissue of origin TCGA analysis methylation regions. In some embodiments, a cancer panel comprises one or more of the regions listed in **Table 8**. For example, a cancer panel comprises at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 8**. In some embodiments, a set of probes are directed to sequences selected from at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 8**.

**Table 8**

| *Closest Gene* | *Chromosome Number* | *Region Start* | *Region Stop* |
|---|---|---|---|
| CLUAP1 | chr16 | 3500621 | 3501549 |
| PPARGC1B | chr5 | 149730215 | 149731334 |
| GCLC | chr6 | 53544322 | 53545301 |
| ZNF648 | chr1 | 182029665 | 182030206 |

| | | | |
|---|---|---|---|
| ITGA6 | chr2 | 172427222 | 172428455 |
| ARHGAP40 | chr20 | 38646114 | 38646491 |
| MARCKS | chr6 | 113858558 | 113859304 |
| PKIG | chr20 | 44531603 | 44532176 |
| G6PC3 | chr17 | 44070557 | 44071133 |
| PPFIBP2 | chr11 | 7513893 | 7514352 |

**[0169]** **Table 9** shows lung methylation regions overlapping in tissue data and TCGA analysis. In some embodiments, a cancer panel comprises one or more of the regions listed in **Table 9**. For example, a cancer panel comprises at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 9**. In some embodiments, a set of probes are directed to sequences selected from at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 9**. These regions may be associated with presence of cancer and are associated with lung tissue, and when combined with the regions in **Table 8**, may be supportive of lung cancer detection.

**Table 9**

| ***Closest Gene*** | ***Chromosome Number*** | ***Region Start*** | ***Region Stop*** |
|---|---|---|---|
| HOPX | chr4 | 56655120 | 56656970 |
| ITGA4 | chr2 | 181456459 | 181458643 |
| PPP1R16B | chr20 | 38804789 | 38807565 |
| HOXA1 | chr7 | 27095185 | 27097167 |
| TCF24 | chr8 | 66961153 | 66963484 |
| LOC100130992 | chr10 | 22252193 | 22254000 |
| LOC100286906 | chr7 | 155371315 | 155374020 |
| EMILIN2 | chr18 | 2847184 | 2848342 |
| IRF4 | chr6 | 391101 | 392638 |
| OR1F2P | chr16 | 3187792 | 3189855 |

**iv. Ovarian Cancer**

**[0170]** **Table 10** shows ovarian tissue of origin TCGA analysis methylation regions. In some embodiments, a cancer panel comprises one or more of the regions listed in **Table 10**. For example, a cancer panel comprises at least one, at least two, at least three, at least four, or all of

the genomic regions listed in **Table 10**. In some embodiments, a set of probes are directed to sequences selected from at least one, at least two, at least three, at least four, or all of the genomic regions listed in **Table 10**.

**Table 10**

| *Closest Gene* | *Chromosome Number* | *Region Start* | *Region Stop* |
|---|---|---|---|
| UBB | chr17 | 16380613 | 16381454 |
| LYPLAL1-DT | chr1 | 219173767 | 219174230 |
| LEMD1 | chr1 | 205430602 | 205431255 |
| SHF | chr15 | 45200050 | 45201385 |
| RSRP1 | chr1 | 25239575 | 25240220 |

**[0171] Table 11** shows ovarian tissue of origin tissue methylation sequencing methylation regions. In some embodiments, a cancer panel comprises one or more of the regions listed in **Table 11**. For example, a cancer panel comprises at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 11**. In some embodiments, a set of probes are directed to sequences selected from at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 11**.

**Table 11**

| *Closest Gene* | *Chromosome Number* | *Region Start* | *Region Stop* |
|---|---|---|---|
| LRRC41 | chr1 | 46302590 | 46304060 |
| LRRC41 | chr1 | 46301753 | 46302464 |
| TMEM101 | chr17 | 44014457 | 44015264 |
| ZNF330 | chr4 | 141220899 | 141221572 |
| RAB35 | chr12 | 120086882 | 120087572 |
| SLC25A29 | chr14 | 100285092 | 100285755 |
| TNK2 | chr3 | 195895389 | 195896162 |
| WBP1 | chr2 | 74457263 | 74457800 |
| NUDT19 | chr19 | 32691703 | 32692417 |
| PALLD | chr4 | 168831678 | 168832743 |

**[0172] Table 12** shows ovarian methylation regions overlapping in tissue data and TCGA analysis. In some embodiments, a cancer panel comprises one or more of the regions listed in

**Table 12**. For example, a cancer panel comprises at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 12**. In some embodiments, a set of probes are directed to sequences selected from at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 12**. These regions may be associated with presence of cancer and may be associated with ovarian tissue, and when combined with the regions in **Table 10** and/or **Table 11**, may be supportive of ovarian cancer detection.

**Table 12**

| *Closest Gene* | *Chromosome Number* | *Region Start* | *Region Stop* |
|---|---|---|---|
| LRRC41 | chr1 | 46302590 | 46304060 |
| TMEM101 | chr17 | 44014457 | 44015264 |
| LRRC41 | chr1 | 46301753 | 46302464 |
| TFAP2E | chr1 | 35576535 | 35578237 |
| ZNF154 | chr19 | 57708612 | 57709440 |
| IRF2BP1 | chr19 | 45876550 | 45877420 |
| LYPLAL1-DT | chr1 | 219173767 | 219174230 |
| SCRG1 | chr4 | 173508370 | 173510040 |
| PDXK | chr21 | 43728573 | 43729711 |
| LOC100286906 | chr7 | 155371315 | 155374020 |

**v. Pancreatic Cancer**

**[0173] Table 13** shows pancreas tissue of origin tissue methylation sequencing methylation regions. In some embodiments, a cancer panel comprises one or more of the regions listed in **Table 13**. For example, a cancer panel comprises at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 13**. In some embodiments, a set of probes are directed to sequences selected from at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 13**.

**Table 13**

| *Closest Gene* | *Chromosome Number* | *Region Start* | *Region Stop* |
|---|---|---|---|
| ZEB2 | chr2 | 144515729 | 144516316 |
| DISP3 | chr1 | 11478576 | 11479792 |
| EMILIN2 | chr18 | 2847184 | 2848342 |

| | | | |
|---|---|---|---|
| CBX8 | chr17 | 79812503 | 79812934 |
| RBFOX3 | chr17 | 79183164 | 79184108 |
| AP3B2 | chr15 | 82680393 | 82680976 |
| KCNA2 | chr1 | 110555424 | 110555713 |
| CTNND2 | chr5 | 11904341 | 11904571 |
| LINC01264 | chr10 | 42932936 | 42933788 |
| HIST1H2BJ | chr6 | 27096849 | 27097190 |

**[0174]** **Table 14** shows pancreas methylation regions overlapping in tissue data and TCGA analysis. In some embodiments, a cancer panel comprises one or more of the regions listed in **Table 14**. For example, a cancer panel comprises at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 14**. In some embodiments, a set of probes are directed to sequences selected from at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 14**. These regions are associated with presence of cancer and are associated with pancreatic tissue, and when combined with the regions in **Table 13**, are supportive of pancreatic cancer detection.

**Table 14**

| *Closest Gene* | *Chromosome Number* | *Region Start* | *Region Stop* |
|---|---|---|---|
| ZNF625 | chr19 | 12155907 | 12156871 |
| LOC392232 | chr8 | 72251140 | 72251956 |
| BASP1-AS1 | chr5 | 17217516 | 17219975 |
| LINC01264 | chr10 | 42932936 | 42933788 |
| KIF7 | chr15 | 89654755 | 89655920 |
| TWIST1 | chr7 | 19112343 | 19112657 |
| SNORD12 | chr20 | 49318252 | 49319561 |
| ARHGEF16 | chr1 | 3393377 | 3394370 |
| SLCO3A1 | chr15 | 91852783 | 91854800 |
| TBX18 | chr6 | 84762839 | 84765672 |

**vi. Prostate Cancer**

**[0175]** **Table 15** lists prostate tissue of origin TCGA analysis methylation regions. In some embodiments, a cancer panel comprises one or more of the regions listed in **Table 15**. For example, a cancer panel comprises at least one, at least two, at least three, at least four, at least

five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 15**. In some embodiments, a set of probes are directed to sequences selected from at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 15**.

**Table 15**

| *Closest Gene* | *Chromosome Number* | *Region Start* | *Region Stop* |
|---|---|---|---|
| SERPINB1 | chr6 | 2841095 | 2842039 |
| FBXO30 | chr6 | 145814298 | 145815785 |
| C2orf88 | chr2 | 190180040 | 190181405 |
| FLOT1 | chr6 | 30742095 | 30744911 |
| KLK10 | chr19 | 51018748 | 51020133 |
| SERPINB9 | chr6 | 2902941 | 2903654 |
| TRIP6 | chr7 | 100866785 | 100867795 |
| ARHGAP40 | chr20 | 38601420 | 38602099 |
| ACSF2 | chr17 | 50425455 | 50426576 |
| KLK13 | chr19 | 51064546 | 51065995 |

**[0176] Table 16** lists prostate tissue of origin tissue methylation sequencing methylation regions. In some embodiments, a cancer panel comprises one or more of the regions listed in **Table 16**. For example, a cancer panel comprises at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 16**. In some embodiments, a set of probes are directed to sequences selected from at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 16**.

**Table 16**

| *Closest Gene* | *Chromosome Number* | *Region Start* | *Region Stop* |
|---|---|---|---|
| MOB3B | chr9 | 27528978 | 27529887 |
| NAV2 | chr11 | 19712785 | 19713147 |
| TSPAN12 | chr7 | 120856772 | 120857377 |
| ABCA1 | chr9 | 104991063 | 104991456 |
| RCCD1 | chr15 | 90954764 | 90955485 |
| C2orf88 | chr2 | 190180174 | 190181293 |
| BMP8B | chr1 | 39789212 | 39789595 |

| | | | |
|---|---|---|---|
| KLF12 | chr13 | 74133823 | 74134156 |
| DMBX1 | chr1 | 46488954 | 46489566 |
| RP9P | chr7 | 32942394 | 32942861 |

**[0177]** **Table 17** lists prostate methylation regions overlapping in tissue data and TCGA analysis. In some embodiments, a cancer panel comprises one or more of the regions listed in **Table 17**. For example, a cancer panel comprises at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 17**. In some embodiments, a set of probes are directed to sequences selected from at least one, at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine, or all of the genomic regions listed in **Table 17**. These regions are associated with presence of cancer and are associated with prostate tissue, and when combined with the regions in **Table 15** and/or **Table 16**, are supportive of prostate cancer detection.

**Table 17**

| ***Closest Gene*** | ***Chromosome Number*** | ***Region Start*** | ***Region Stop*** |
|---|---|---|---|
| MOB3B | chr9 | 27528978 | 27529887 |
| IRF2BP1 | chr19 | 45876550 | 45877420 |
| AKR1B1 | chr7 | 134458200 | 134459533 |
| KLK10 | chr19 | 51018748 | 51020133 |
| ZSCAN12 | chr6 | 28399347 | 28400210 |
| RNVU1-8 | chr1 | 147083835 | 147085351 |
| PPP1R16B | chr20 | 38804789 | 38807565 |
| LOC100130298 | chr8 | 60909799 | 60910469 |
| LOC100507557 | chr6 | 145814298 | 145815785 |
| UBXN10 | chr1 | 20185867 | 20186730 |

**[0178]** In an aspect, the present disclosure provides a method for identifying a methylation signature indicative of a biological characteristic, the method comprising: obtaining data for a population comprising a plurality of genomic methylation data sets associated with cell proliferative disorder status, wherein each of said genomic methylation data sets are associated with biological information for a corresponding sample; segregating the methylation data sets into a first group corresponding to one tissue or cell type possessing the biological characteristic and a second group corresponding to a plurality of tissue or cell types not possessing the

biological characteristic; matching methylation data from the first group to methylation data from the second group on a site-by-site basis across the genome; identifying a set of CpG sites on a site-by-site basis across the genome that meet a predetermined threshold for establishing differential methylation between the first and second groups; identifying, using the set of CpG sites, target genomic regions comprising at least one, at least two, at least three, or more than three differentially methylated CpGs within about 30 to 300 bp that meet said predetermined criteria, to identify differentially methylated genomic regions that provide the methylation signature indicative of the biological characteristic associated with the presence of a cell proliferative disorder.

**[0179]** In some examples, the target genomic region comprises at least one, at least two, at least three, or more than three differentially methylated CpG sites within a region having a length of about 30 to 150 bp, about 40 to 150 bp, about 50 to 150 bp, about 75 to 150 bp, about 100 to 150 bp, about 150 to 300 bp, about 150 to 250 bp, about 150 to 200 bp, about 200 to 300 bp, or about 250 to 300 bp.

**[0180]** In some examples, the target genomic region comprises at least four differentially methylated CpG sites, at least five differentially methylated CpG sites, at least six differentially methylated CpG sites, at least seven differentially methylated CpG sites, at least eight differentially methylated CpG sites, at least nine differentially methylated CpG sites, at least ten differentially methylated CpG sites, at least 12 differentially methylated CpG sites, or at least 15 differentially methylated CpG sites.

**[0181]** In some embodiments, the method further comprises validating the extended target genomic regions by testing for differential methylation within the extended target genomic regions using DNA from at least one independent sample possessing the biological trait and DNA from at least one independent sample not possessing the biological sample.

**[0182]** In some embodiments, the identifying further comprises limiting the set of CpG sites to CpG sites that further exhibit differential methylation with peripheral blood mononuclear cells from a control sample.

**[0183]** In some embodiments, the predetermined threshold is at least about 50% methylation in the first group.

**[0184]** In some embodiments, the predetermined threshold is a difference in average methylation between the first and second groups of at least about 0.3.

**[0185]** In some embodiments, the biological trait comprises malignancy.

**[0186]** In some embodiments, the biological trait comprises a cancer type.

**[0187]** In some embodiments, the biological trait comprises a cancer stage.

**[0188]** In some embodiments, the biological trait comprises a cancer classification.

**[0189]** In some embodiments, the cancer classification comprises a cancer grade.

**[0190]** In some embodiments, the cancer classification comprises a histological classification.

**[0191]** In some embodiments, the biological trait comprises a metabolic profile.

**[0192]** In some embodiments, the biological trait comprises a mutation.

**[0193]** In some embodiments, the mutation is a disease-associated mutation.

**[0194]** In some embodiments, the biological trait comprises a clinical outcome.

**[0195]** In some embodiments, the biological trait comprises a drug response.

**[0196]** In some embodiments, the method further comprises designing a plurality of PCR primers pairs to amplify portions of the extended target genomic regions, each of the portions comprising at least one differentially methylated CpG site.

**[0197]** In some embodiments, the designing of the plurality of primer pairs comprising converting non-methylated cytosines uracil, to simulate cytosine to uracil conversion, and designing the primer pairs using the converted sequence.

**[0198]** In some embodiments, the primer pairs are designed to have a methylation bias.

**[0199]** In some embodiments, the primer pairs are methylation-specific.

**[0200]** In some embodiments, the primer pairs have no CpG residues within them having no preference for methylation status.

**[0201]** In an aspect, the present disclosure provides a method for synthesizing primer pairs specific to a methylation signature, the method comprising: carrying out a method of the present disclosure, and synthesizing the designed primer pairs.

## IV. NUCLEIC ACID CONVERSION AND METHYLATION SEQUENCING

### A. Nucleic Acid Treatment

**[0202]** Various methods are available for methylation sequencing that include chemical-based and enzymatic-based conversion of nucleic acid bases to distinguish methylated from unmethylated cytosines in a nucleic acid sequence. These assays allow for determination of the methylation state of one or a plurality of CpG dinucleotides (e.g., CpG islands) within a DNA sequence. Such assays may comprise, among other techniques, DNA sequencing of bisulfite-treated DNA, or enzymatic-treated DNA, polymerase chain reaction (PCR) (for sequence-specific amplification), quantitative PCR (qPCR), or digital droplet PCR (ddPCR), Southern blot analysis. In various examples, DNA in a biological sample is treated in such a manner that cytosine bases which are unmethylated at the 5'-position are converted to uracil, thymine, or another base which is dissimilar to cytosine in terms of hybridization behavior. This process may be referred to as “conversion”.

**[0203]** In some embodiments, a reagent converts cytosine bases which are unmethylated at the 5'-position to uracil, thymine, or another base which is dissimilar to cytosine in terms of hybridization behavior.

**[0204]** Bisulfite modification of DNA generally refers to a tool used to assess CpG methylation status. A method for analyzing DNA for the presence of 5-methylcytosine may be based upon the reaction of bisulfite with cytosine whereby, upon subsequent alkaline desulfonation, cytosine is converted to uracil which corresponds to thymine with respect to base pairing behavior. For example, genomic sequencing may be adapted for analysis of DNA methylation patterns and 5-methylcytosine distribution by using bisulfite treatment (e.g., as described by Frommer et al., Proc. Natl. Acad. Sci. USA 89:1827-1831, 1992, the contents of which are incorporated herein by reference). Significantly, however, 5-methylcytosine may remain unmodified under these conditions. Consequently, the original DNA may be converted in such a manner that methylcytosine, which originally could not be distinguished from cytosine by hybridization behavior, can now be detected as the only remaining cytosine using various molecular biological techniques, for example, by amplification and hybridization, or by sequencing. In various examples, other reagents may affect the same result as bisulfite modification useful for methylation sequencing.

**[0205]** A direct sequencing method may employ bisulfite-treated DNA amplified with PCR useful with whole-genome bisulfite sequencing (WGBS) or targeted bisulfite sequencing.

**[0206]** Targeted Bisulfite Sequencing is a commercially available NGS method used to evaluate site-specific DNA methylation changes. Probes may be designed to be strand-specific as well as bisulfite-specific. Both methylated and unmethylated sequences may be amplified. The process may be similar to pyrosequencing, but may offer a much higher throughput overall. In some embodiments, next-generation sequencing platforms are used to deliver large amounts of useful DNA methylation information (e.g., EPIGENTEK, Farmingdale, NY and ZYMO RESEARCH, Irvine, CA). The methylation analysis at single-base resolution of individual cytosine in DNA may be facilitated by bisulfite treatment of DNA followed by PCR amplification of targeted region, library construction, and sequencing of the amplicon regions. Specific primers may be designed for the region of interest and cytosine methylation changes may be evaluated within that region. Each DNA methylation site of interest may be assessed at high-sequencing depth of coverage for accurate, quantitative, and single-base resolution data output.

**[0207]** Enzymatic methyl sequencing (EM-seq) may rely on enzymatic conversion of nucleic acids for methylome analysis. The process of generating EM-seq libraries may not damage DNA in the same way as bisulfite sequencing. EM-seq libraries may give higher PCR yields despite using fewer PCR cycles for all DNA input amounts, indicating that less DNA is lost during

enzymatic treatment and library preparation, as compared to whole genome bisulfite sequencing (WGBS). Reduced PCR cycles, in turn, may translate into more complex libraries and fewer PCR duplicates during sequencing. EM-seq libraries also may have larger average insert sizes than WGBS which further supports the fact that DNA remains intact. In the EM-seq workflow, TET2 oxidizes 5-mC and 5-hmC, providing protection from deamination by APOBEC in the next operation. In contrast, unmodified cytosines may be deaminated to uracils. In some embodiments, the targeted method comprises enzymatic conversion of nucleic acid (TEM-seq). In some embodiments, the methylation sequencing methods may be accomplished with the NEBNEXT® Enzymatic Methyl-seq (New England Biolabs, Ipswich, MA) which may be useful for identification of 5-mC and 5-hmC.

**[0208]** In another example, 5-hmC may be also detected using TET-assisted bisulfite sequencing (TAB-seq) (e.g., as described by Yu, M., et al. (2012). Nat. Protoc. 7, 2159-2170, the contents of which are incorporated herein by reference) (WiseGene; Illumina). Fragmented DNA may be enzymatically modified using sequential T4 Phage β-glucosyltransferase (T4-BGT), and then Ten-eleven translocation (TET) dioxygenase treatments before the addition of sodium bisulfite. T4-BGT is used to glucosylate 5-hmC to form β-glucosyl-5-hydroxymethylcytosine (5-ghmC) and TET is then used to oxidize 5-mC to 5-caC. Only 5-ghmC is protected from subsequent deamination by sodium bisulfite and this enables 5-hmC to be distinguished from 5-mC by sequencing.

**[0209]** Oxidative bisulfite sequencing (oxBS) provides another method to distinguish between 5-mC and 5-hmC (e.g., as described by Booth, M.J., et al., 2012 Science 336: 934-937, the contents of which are incorporated herein by reference). The oxidation reagent potassium perruthenate converts 5-hmC to 5-formylcytosine (5-fC) and subsequent sodium bisulfite treatment deaminates 5-fC to uracil. 5-mC remains unchanged and can therefore be identified using this method.

**[0210]** APOBEC-coupled epigenetic sequencing (ACE-seq) excludes bisulfite conversion altogether and relies on enzymatic conversion to detect 5-hmC (e.g., as described by Schutsky, E.K., et al., Nat. Biotechnol., 2018 Oct 8, the contents of which are incorporated herein by reference). With this method, T4-BGT glucosylates 5-hmC to 5-ghmC, which protects 5-hmC from deamination by Apolipoprotein B mRNA editing enzyme subunit 3A (APOBEC3A). Cytosine. 5-mC is deaminated by APOBEC3A and sequenced as thymine.

**[0211]** In another example, a bisulfite-free and base-level-resolution sequencing method, TET-assisted pyridine borane sequencing (TAPS), may be used for detection of 5-mC and 5-hmC. TAPS combines ten-eleven translocation (TET) oxidation of 5-mC and 5-hmC to 5-carboxylcytosine (5-caC) with pyridine borane reduction of 5-caC to dihydrouracil (DHU).

Subsequent PCR converts DHU to thymine, enabling a C-to-T transition of 5-mC and 5-hmC. TAPS detects modifications directly with high sensitivity and specificity, without affecting unmodified cytosines (e.g., as described by Liu, Y., et al. Nat Biotechnol. 2019 Apr;37(4):424-429, the contents of which are incorporated herein by reference).

**[0212]** TET-assisted 5-methylcytosine sequencing (TAmC-seq) enriches for 5-mC loci and utilizes two sequential enzymatic reactions followed by an affinity pull-down (Zhang, L. 2013, Nat Commun 4: 1517). Fragmented DNA is treated with T4-BGT which protects 5-hmC by glucosylation. The enzyme mTET1 is then used to oxidize 5-mC to 5-hmC, and T4-BGT labels the newly formed 5-hmC using a modified glucose moiety (6-N3-glucose). Click chemistry may be used to introduce a biotin tag which enables enrichment of 5-mC-containing DNA fragments for detection and genome wide profiling.

**B. Next-generation Sequencing**

**[0213]** In some embodiments, the generating of sequencing reads is carried out by next-generation sequencing (NGS). NGS may permit a high depth of reads to be achieved for a given region. Such high-throughput methods include, for example, Illumina (Solexa) sequencing, DNB-Sequencer T7 or G400 (MGI Tech Co., Ltd), GenapSys sequencing (GenapSys, Inc.), Roche 454 sequencing (Roche Sequencing Solutions, Inc.), Ion Torrent sequencing (Thermo Fisher Scientific), and SOLiD sequencing (Thermo Fisher Scientific). The number of sequencing reads may be adjusted depending on DNA input amount and depth of data required for analysis.

**[0214]** In some embodiments, the generating of sequencing reads is carried out simultaneously for samples obtained from multiple patients, wherein the cell-free nucleic acid fragments are barcoded for each patient. Simultaneous generation of sequencing reads permits parallel analysis of a plurality of patients in one sequencing run.

**[0215]** In another aspect, the present disclosure provides a kit for detecting a tumor comprising reagents for carrying out the aforementioned method and instructions for detecting the tumor signals. Reagents may include, for example, primer sets, PCR reaction components, and/or sequencing reagents.

**C. Targeted Sequencing**

**[0216]** In targeted methylation sequencing approaches, targeted regions in a biological sample such as cfDNA may be analyzed to determine the methylation state of the target gene sequences. In some embodiments, the target region comprises, or hybridizes under stringent conditions to, contiguous nucleotides of target regions of interest, such as at least about 16 contiguous

nucleotides of a target region of interest. In different examples, targeted sequencing may be accomplished using hybridization capture and amplicon sequencing approaches.

**D. Hybridization Capture**

**[0217]** The hybridization method provided herein may be used in various formats of nucleic acid hybridizations, such as in-solution hybridization and such as hybridization on a solid support (e.g., Northern, Southern, and *in situ* hybridization on membranes, microarrays, and cell/tissue slides). In particular, the method is suitable for in-solution hybrid capture for target enrichment of certain types of genomic DNA sequences (e.g., exons) employed in targeted next-generation sequencing. For hybrid capture approaches, a cell-free nucleic acid sample may be subjected to library preparation. As used herein, "library preparation" comprises end-repair, A-tailing, adapter ligation, or any other preparation performed on the cell-free DNA to permit subsequent sequencing of DNA. In certain examples, a prepared cell-free nucleic acid library sequence contains adapters, sequence tags, or index barcodes that are ligated onto cell-free nucleic acid sample molecules. Various commercially available kits may be used to facilitate library preparation for next-generation sequencing approaches. Next-generation sequencing library construction may comprise preparing nucleic acids targets using a coordinated series of enzymatic reactions to produce a random collection of DNA fragments of specific size for high throughput sequencing. Advances and the development of various library preparation technologies have expanded the application of next-generation sequencing to fields such as transcriptomics and epigenetics.

**[0218]** Improvements in sequencing technologies have resulted in changes and improvements to library preparation. Next-generation sequencing library preparation kits, developed by companies such as Agilent, Bioo Scientific, Kapa Biosystems, New England Biolabs, Illumina, Life Technologies, Pacific Biosciences, and Roche may provide consistency and reproducibility to various molecular biology reactions that ensure compatibility with the latest NGS instrument technology.

**[0219]** In various examples for targeted capture gene panels, various library preparation kits may be selected from Nextera Flex (Illumina), IonAmpliseq (Thermo Fisher Scientific), Genexus (Thermo Fisher Scientific), Agilent ClearSeq (Illumina), Agilent SureSelect Capture (Illumina), Archer FusionPlex (Illumina), BiooScientific NEXTflex (Illumina), IDT xGen (Illumina), Illumina TruSight (Illumina), Nimblegene SeqCap (Illumina), and Qiagen GeneRead (Illumina).

**[0220]** In some embodiments, the hybrid capture method is carried out on the prepared library sequences using specific probes. In some embodiments, the term "specific probe", as used

herein, generally refers to a probe that is specific for known methylation sites. In some embodiments, the specific probes are designed based on using human genome as a reference sequence and using specified genomic regions known to have methylation sites as target sequences. Specifically, genomic regions known to have methylation sites may comprise at least one of the following: a promoter region, a CpG island region, a CGI shore region, and an imprinted gene region. Therefore, when carrying out the hybrid capture by using the specific probes of some embodiments, the sequences in the sample genome which are complimentary to the target sequences, e.g., regions in the sample genome known to have methylation sites (which are also referred to as "specified genomic regions" herein) may be captured efficiently.

**[0221]** In some embodiments, the methylated regions described herein are used for designing the specific probes. In some embodiments, the specific probes are designed using commercially available methods such as for example an eArray system. The length of the probes may be sufficient to hybridize with sufficient specificity to the methylated region of interest. In various examples, the probe is a 10-mer, 11-mer, 12-mer, 13-mer, 14-mer 15-mer, 16-mer, 17-mer, 18-mer, 19-mer, or 20-mer.

**[0222]** The regions listed in **Tables 1-17** may be screened using database resources (such as gene ontology). According to the principle of complementary base pairing, a single-stranded capture probe may be combined with a single-stranded target sequence complementarily, so as to capture the target region successfully. In some embodiments, the designed probes may be designed as a solid capture chip (wherein the probes are immobilized on a solid support) or be designed as a liquid capture chip (wherein the probes are free in the liquid), however, limited by various factors, such as probe length, probe density, and high cost etc. The solid capture chip is rarely used, while the liquid capture chip is used more frequently.

**[0223]** In some embodiments, compared with normal sequences (where the average content of A, T, C, and G base is 25% each, respectively), GC-rich sequences (where the content of GC bases is higher than 60%) in nucleic acid may lead to the reduction of capture efficiency because of the molecular structure of C and G bases. For the key research regions, for example, CGI regions (CpG islands), designing an increased amount of the probes to obtain sufficient and accurate CGI data may be recommended.

### E. Amplicon-Based Sequencing

**[0224]** Fragments of the converted DNA may be amplified. In some embodiments, the amplifying is carried out with primers designed to anneal to methylation converted target sequences having at least one methylated site therein. Methylation sequencing conversion results in unmethylated cytosines being converted to uracil, while 5-methylcytosine is unaffected.

"Converted target sequences" may thus be understood as sequences in which cytosines known to be methylation sites are fixed as "C" (cytosine), while cytosines known to be unmethylated may be fixed as "U" (uracil; which may be treated as "T" (thymine) for primer design purposes).

**[0225]** In various examples, the source of the DNA may be cell-free DNA from whole blood, plasma, serum, or genomic DNA extracted from cells or tissue. In some embodiments, the size of the amplified fragment is between about 100 and 200 base pairs in length. In some embodiments, the DNA source is extracted from cellular sources (e.g., tissues, biopsies, or cell lines), and the amplified fragment is between about 100 and 350 base pairs in length. In some embodiments, the amplified fragment comprises at least one 20 base pair sequence comprising at least one, at least two, at least three, or more than three CpG dinucleotides. The amplification may be carried out using sets of primer oligonucleotides according to the present disclosure, and may use a heat-stable polymerase. The amplification of several DNA segments may be carried out simultaneously in one and the same reaction vessel. In some embodiments of the method, two or more fragments are amplified simultaneously. For example, the amplification may be carried out using a polymerase chain reaction (PCR).

**[0226]** Primers designed to target such sequences may exhibit a degree of bias towards converted methylated sequences. In some embodiments, the PCR primers are designed to be methylation specific for targeted methylation-sequencing applications, which may allow for greater sensitivity in some applications. For instance, primers may be designed to include a discriminatory nucleotide (specific to a methylated sequence following bisulfite conversion) positioned to achieve optimal discrimination, e.g., in PCR applications. The discriminatory may be positioned at the 3' ultimate or penultimate position.

**[0227]** Primers may be designed to amplify DNA fragments based on the general size range for circulating DNA. Optimizing primer design to take into account target size may increase the sensitivity of the method according to this example. In some embodiments, the primers are designed to amplify DNA fragments 75 to 350 bp in length. The primers may be designed to amplify regions that are about 50 to 200, about 75 to 150, or about 100 or 125 bp in length.

**[0228]** In some embodiments of the method, the methylation status of preselected CpG positions within the nucleic acid sequences may be detected by the amplicon-based approach using of methylation-specific primer oligonucleotides. The use of methylation status specific primers for the amplification of bisulfite treated DNA may allow the differentiation between methylated and unmethylated nucleic acids. MSP primers pairs may contain at least one primer which hybridizes to a converted CpG dinucleotide. Therefore, the sequence of said primers may comprises at least one CpG, TpG, or CpA dinucleotide. MSP primers specific for non-methylated DNA may contain a "T" at the 3' position of the C position in the CpG. Therefore,

the base sequence of said primers may be required to comprise a sequence having a length of at least 18 nucleotides which hybridizes to a pretreated nucleic acid sequence and sequences complementary thereto, wherein the base sequence of said oligomers comprises at least one CpG, TpG, or CpA dinucleotide. In some embodiments of the method, the MSP primers comprise between 2 and 5 CpG, TpG, or CpA dinucleotides. In some embodiments, the dinucleotides are located within the 3' half of the primer, e.g., wherein a primer is 18 bases in length the specified dinucleotides are located within the first 9 bases form the 3' end of the molecule. In addition to the CpG, TpG, or CpA dinucleotides, the primers may further comprise several methyl converted bases (e.g., cytosine converted to thymine, or on the hybridizing strand, guanine converted to adenosine). In some embodiments, the primers are designed so as to comprise no more than 2 cytosine or guanine bases.

**[0229]** In some embodiments, each of the regions is amplified in sections using multiple primer pairs. In some embodiments, these sections are non-overlapping. The sections may be immediately adjacent or spaced apart (e.g., spaced apart up to 10, 20, 30, 40, or 50 bp). Since target regions (including CpG islands, CpG shores, and/or CpG shelves) are usually longer than 75 to 150 bp, this example may permit the methylation status of sites across more (or all) of a given target region to be assessed.

**[0230]** Primers may be designed for target regions using suitable tools such as Primer3, Primer3Plus, Primer-BLAST, etc. As discussed, bisulfite conversion results in cytosine converting to uracil and 5'-methyl-cytosine converting to thymine. Thus, primer positioning or targeting may make use of bisulfite converted methylated sequences, depending on the degree of methylation specificity required.

**[0231]** Target regions for amplification may be designed to have at least 10 CpG dinucleotide methylation sites. In some examples, however, amplification of regions having more than 10 CpG methylation site may be advantageous. For instance, a sequence read 300 bp long may have about 10, 20, 30, 40, or 50 CpG methylation sites that are methylated in a nucleic acid sample associated with a cell proliferative disorder. In various examples, the methylation regions identified in **Tables 1-17** may have 25, 50, 100, 200, 300, 400, or 500 CpG methylation sites that are methylated in a nucleic acid sample associated with a cell proliferative disorder. In some embodiments, the primers are designed to amplify DNA fragments comprising 3 to 20 CpG methylation sites in a targeted region. Overall, this approach may permit a larger number of methylation sites to be queried within a single sequencing read and may provide additional certainty (exclusion of false positives) because multiple concordant methylations may be detected within a single sequencing read. In some embodiments, the tumor signals comprise more than two methylated regions selected from **Tables 1-17**. Detection of multiple tumor

signals, in this example, may increase confidence in tumor detection. Such signals may be at the same or at different sites. In some embodiments, the detection of more than one of the tumor signals at the same region is indicative of a tumor.

**[0232]** In some embodiments, the number of CpG sites in an identified methylated region may be modeled between two populations having a different characteristic of a cell proliferative disorder to identify a methylation threshold where the number of CpG sites in a region that exceeds the threshold is indicative of a cell proliferative disorder.

**[0233]** In various examples, the number of CpG sites in an identified methylated region that indicates cancer is 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, or 18, where the presence of methylated CpGs that exceeds this identified number is indicative of cancer and may be used as an input feature into a machine learning model used as a classifier to stratify a population into healthy individuals and those having cancer.

**[0234]** Detection of multiple tumor signals indicative of methylation at the same site in the genome, in this example, may increase confidence in tumor detection. Detection of methylation at adjacent sites in the genome may also increase confidence in tumor detection, even if the signals are derived from different sequencing reads. Detection of methylation at adjacent sites in the genome reflects another type of signal concordance. In some embodiments, the detection of adjacent or overlapping tumor signals across at least two different sequencing reads is indicative of a tumor. In some embodiments, the adjacent or overlapping tumor signals are within the same CpG island. In some embodiments, the detection of 3 to 34 proximal methylated sites in a cell-free DNA fragment is indicative of a tumor. In some embodiments, the detection of 3 to 34 methylated CpG sites in a fragment is used to identify a threshold to distinguish between a population of individuals having a characteristic (e.g., healthy, disease, or stage of disease). In some embodiments, the detection of about 4 to 10, about 4 to 15, about 10 to 20, about 15 to 20, about 15 to 25, about 20 to 25, about 20 to 34, about 25 to 34, or about 30 to 34 methylated proximal CpG sites in a read fragment is used to identify a threshold to distinguish between a population of individuals having a characteristic (e.g., healthy, disease, or stage of disease). As used herein, the term “proximal CpG site” refers to CpG sites that are adjacent or within about 2 to 10 CpG sites of each other and where the CpG sites on the same nucleic acid fragment in a cell-free nucleic acid sample.

**[0235]** In some embodiments, the amplification is carried out with more than 100 primer pairs. The amplification may be carried out with about 10, about 20, about 30, about 40, about 50, about 60, about 70, about 80, about 90, about 100, about 110, about 120, about 130, about 140, about 150, or more primer pairs. In some embodiments, the amplification is a multiplex amplification. Multiplex amplification permits large amount of methylation information to be

gathered from many target regions in the genome in parallel, even from cfDNA samples in which DNA is generally not plentiful. The multiplexing may be scaled up to a platform such as ION AmpliSeq, in which, e.g., up to about 24,000 amplicons may be queried simultaneously. In some embodiments, the amplification is nested amplification. A nested amplification may improve sensitivity and specificity.

**[0236]** Further, another rapid and robust protocol for the parallel examination of multiple methylated sequences termed simultaneous targeted methylation sequencing (sTM-Seq). Key features of this technique include the elimination of the need for large amounts of high-molecular weight DNA and the nucleotide specific distinction of both 5-methylcytosine (5-mC) and 5-hydroxymethylcytosine (5-hmC). Moreover, sTM-Seq may be scalable and may be used to investigate multiple loci in dozens of samples within a single sequencing run. Freely available web-based software and universal primers for multipurpose barcoding, library preparation, and customized sequencing make sTM-Seq affordable, efficient, and widely applicable. (as described by Asmus, N. et al., Curr Protoc Hum Genet. 2019 Apr;101(1), the contents of which are incorporated herein by reference).

**[0237]** Generally, the methods and systems provided herein may be useful for preparation of cell-free polynucleotide sequences to a downstream application sequencing reaction. In some embodiments, a sequencing method is classic Sanger sequencing. Sequencing methods may include, but are not limited to: high-throughput sequencing, pyrosequencing, sequencing-by-synthesis, single-molecule sequencing, nanopore sequencing, semiconductor sequencing, sequencing-by-ligation, sequencing-by-hybridization, RNA-Seq (Illumina), Digital Gene Expression (Helicos), Next-generation sequencing, Single Molecule Sequencing by Synthesis (SMSS)(Helicos), massively-parallel sequencing, Clonal Single Molecule Array (Solexa), shotgun sequencing, Maxim-Gilbert sequencing, primer walking, and any other sequencing methods.

**[0238]** Pyrosequencing is a real-time sequencing technology based on luminometric detection of pyrophosphate release upon nucleotide incorporation which is suitable for simultaneous analysis and quantification of the methylation degree of several CpG positions. After conversion of genomic DNA, a region of interest may be amplified by polymerase chain reaction (PCR) with one of the two primers being biotinylated. The PCR-generated template may be rendered single stranded and a Pyrosequencing primer is annealed to analyze quantitatively CpG positions. After bisulfite treatment and PCR, the degree of each methylation at each CpG position in a sequence may be determined from the ratio of T and C signals reflecting the proportion of unmethylated and methylated cytosines at each CpG site in the original sequence.

## V. CLASSIFIERS, MACHINE LEARNING MODELS & SYSTEMS

**[0239]** In various examples, methylation sequencing features may be used as input datasets into trained algorithms (e.g., machine learning models or classifiers) to identify correlations between sequence composition and patient groups. Examples of such patient groups include presence of diseases or conditions, stages, subtypes, responders vs. non-responders, and progressors vs. non-progressors. In various examples, feature matrices may be generated to compare samples obtained from individuals with known conditions or characteristics. In some embodiments, samples may be obtained from healthy individuals, or individuals who do not have any of the known indications and samples from patients known to have cancer.

**[0240]** As used herein, relating to machine learning and pattern recognition, the term "feature" generally refers to an individual measurable property or characteristic of a phenomenon being observed. The concept of "feature" may be related to that of explanatory variable used in statistical techniques such as for example, but not limited to, linear regression and logistic regression. Features may numeric, but structural features such as strings and graphs may be used in syntactic pattern recognition.

**[0241]** The term "input features" (or "features"), as used herein, generally refers to variables that are used by the trained algorithm (e.g., model or classifier) to predict an output classification (label) of a sample, e.g., a condition, sequence content (e.g., mutations), suggested data collection operations, or suggested treatments. Values of the variables may be determined for a sample and used to determine a classification.

**[0242]** In various examples, input features of genetic data may include: aligned variables that relate to alignment of sequence data (e.g., sequence reads) to a genome and non-aligned variables, e.g., that relate to the sequence content of a sequence read, a measurement of protein or autoantibody, or the mean methylation level at a genomic region. Input features may be genetic features such as, chromatin accessibility (for example transcription factor binding features), nucleosome positioning features (for example V-plot measures and cfDNA measurement over a transcription start site), or cell type deconvolution (for example FREE-C deconvolution). Metrics that may be used in methylation analysis include, but are not limited to, base wise methylation percent for CpG, CHG, CHH, conversion efficiency (100-Mean methylation percent for CHH), hypomethylated blocks, methylation levels (global mean methylation for CPG, CHH, CHG, fragment length, fragment midpoint, number of methylated CpGs per fragment, fraction of CpG methylation to total CpG per fragment, fraction of CpG methylation to total CpG per region, fraction of CpG methylation to total CpG in panel, dinucleotide coverage (normalized coverage of di-nucleotide), evenness of coverage (unique CpG sites at lx and 10x mean genomic coverage (for S4 runs), mean CpG coverage (depth)

globally, and mean coverage at CpG islands, CGI shelves, or CGI shores. These metrics may be used as feature inputs for machine learning methods and models.

**[0243]** For a plurality of assays, the system may identify feature sets to be analyzed using a trained algorithm (e.g., machine learning model or classifier). The system may perform an assay on each molecule class and forms a feature vector from the measured values. The system may analyze the feature vector using the machine learning model and obtain an output classification of whether the biological sample has a specified property.

**[0244]** In some embodiments, the machine learning model outputs a classifier capable of distinguishing between two or more groups or classes of individuals or features in a population of individuals or features of the population. In some embodiments, the classifier is a trained machine learning classifier.

**[0245]** In some embodiments, the informative loci or features of biomarkers in a cancer tissue are assayed to form a profile. Receiver-operating characteristic (ROC) curves may be generated by plotting the performance of a particular feature (e.g., any of the biomarkers described herein and/or any item of additional biomedical information) in distinguishing between two populations (e.g., individuals responding and not responding to a therapeutic agent). In some embodiments, the feature data across the entire population (e.g., the cases and controls) are sorted in ascending order based on the value of a single feature.

**[0246]** In various examples, the specified property is selected from healthy vs. cancer, disease subtype, disease stage, progressor vs. non-progressor, and responder vs. non-responder.

**A. Data Analysis**

**[0247]** In some examples, the present disclosure provides a system, method, or kit having data analysis realized in software application, computing hardware, or both. In various examples, the analysis application or system comprises at least a data receiving module, a data pre-processing module, a data analysis module (which can operate on one or more types of genomic data), a data interpretation module, or a data visualization module. In some embodiments, the data receiving module can comprise computer systems that connect laboratory hardware or instrumentation with computer systems that process laboratory data. In some embodiments, the data pre-processing module comprise hardware systems or computer software that performs operations on the data in preparation for analysis. Examples of operations that may be applied to the data in the pre-processing module include affine transformations, denoising operations, data cleaning, reformatting, or subsampling. A data analysis module, which may be specialized for analyzing genomic data from one or more genomic materials, may, for example, perform probabilistic and statistical analysis on assembled genomic sequences to identify abnormal

patterns related to a disease, pathology, state, risk, condition, or phenotype. A data interpretation module may use analysis methods, for example, drawn from statistics, mathematics, or biology, to support understanding of the relation between the identified abnormal patterns and health conditions, functional states, prognoses, or risks. A data visualization module may use methods of mathematical modeling, computer graphics, or rendering to create visual representations of data that can facilitate the understanding or interpretation of results.

**[0248]** In various examples, machine learning methods may be applied to distinguish samples in a population of samples. In some embodiments, machine learning methods are applied to distinguish samples between healthy and advanced disease (e.g., adenoma) samples.

**[0249]** In some embodiments, the one or more machine learning operations used to train the prediction engine are selected from the group consisting of: a generalized linear model, a generalized additive model, a non-parametric regression operation, a random forest classifier, a spatial regression operation, a Bayesian regression model, a time series analysis, a Bayesian network, a Gaussian network, a decision tree learning operation, an artificial neural network, a recurrent neural network, a convolutional neural network, a reinforcement learning operation, linear or non-linear regression operations, a support vector machine, a clustering operation, and a genetic algorithm operation.

**[0250]** In various examples, computer processing methods are selected from the group consisting of logistic regression, multiple linear regression (MLR), dimension reduction, partial least squares (PLS) regression, principal component regression, autoencoders, variational autoencoders, singular value decomposition, Fourier bases, wavelets, discriminant analysis, support vector machine, decision tree, classification and regression trees (CART), tree-based methods, random forest, gradient boost tree, logistic regression, matrix factorization, multidimensional scaling (MDS), dimensionality reduction methods, t-distributed stochastic neighbor embedding (t-SNE), multilayer perceptron (MLP), network clustering, neuro-fuzzy, and artificial neural networks.

**[0251]** In some examples, the methods disclosed herein can include computational analysis on nucleic acid sequencing data of samples from an individual or from a plurality of individuals.

### B. Classifier Generation

**[0252]** In an aspect, the disclosed systems and methods provide a classifier generated based on feature information derived from methylation sequence analysis from biological samples of cfDNA. The classifier may form part of a predictive engine for distinguishing groups in a population based on sequence features identified in biological samples such as cfDNA.

**[0253]** In some embodiments, a classifier is created by normalizing the sequence information by formatting similar portions of the sequence information into a unified format and a unified scale; storing the normalized sequence information in a columnar database; training a prediction engine by applying one or more one machine learning operations to the stored normalized sequence information, the prediction engine mapping, for a particular population, a combination of one or more features; applying the prediction engine to the accessed field information to identify an individual associated with a group; and classifying the individual into a group.

**[0254]** In some embodiments, a hierarchy is created by normalizing the sequence information by formatting similar portions of the sequence information into a unified format and a unified scale; storing the normalized sequence information in a columnar database; training a prediction engine by applying one or more one machine learning operations to the stored normalized sequence information, the prediction engine mapping, for a particular population, a combination of one or more features; applying the prediction engine to the accessed field information to identify an individual associated with a group; and classifying the individual into a group.

**[0255]** Specificity, as used herein, generally refers to "the probability of a negative test among those who are free from the disease". Specificity may be calculated by the number of disease-free persons who tested negative divided by the total number of disease-free individuals.

**[0256]** In various examples, the model, classifier, or predictive test has a specificity of at least 40%, at least 45%, at least 50%, at least 55%, at least 60%, at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 95%, or at least 99%.

**[0257]** Sensitivity, as used herein, generally refers to "the probability of a positive test among those who have the disease". Sensitivity may be calculated by the number of diseased individuals who tested positive divided by the total number of diseased individuals.

**[0258]** In various examples, the model, classifier, or predictive test has a sensitivity of at least 40%, at least 45%, at least 50%, at least 55%, at least 60%, at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 95%, or at least 99%.

### C. Digital Processing Device

**[0259]** In some examples, the subject matter described herein may include a digital processing device or use of the same. In some examples, the digital processing device may include one or more hardware central processing units (CPU), graphics processing units (GPU), or tensor processing units (TPU) that carry out the device's functions. In some examples, the digital processing device may include an operating system configured to perform executable instructions.

**[0260]** In some examples, the digital processing device may optionally be connected a computer network. In some examples, the digital processing device may be optionally connected to the Internet. In some examples, the digital processing device may be optionally connected to a cloud computing infrastructure. In some examples, the digital processing device may be optionally connected to an intranet. In some examples, the digital processing device may be optionally connected to a data storage device.

**[0261]** Non-limiting examples of suitable digital processing devices include server computers, desktop computers, laptop computers, notebook computers, sub-notebook computers, netbook computers, netpad computers, set-top computers, handheld computers, Internet appliances, mobile smartphones, and tablet computers. Suitable tablet computers may include, for example, those with booklet, slate, and convertible configurations.

**[0262]** In some examples, the digital processing device may include an operating system configured to perform executable instructions. For example, the operating system may include software, including programs and data, which manages the device's hardware and provides services for execution of applications. Non-limiting examples of operating systems include Ubuntu, FreeBSD, OpenBSD, NetBSD®, Linux, Apple® Mac OS X Server®, Oracle® Solaris®, Windows Server®, and Novell® NetWare®. Non-limiting examples of suitable personal computer operating systems include Microsoft® Windows®, Apple® Mac OS X®, UNIX®, and UNIX-like operating systems such as GNU/Linux®. In some examples, the operating system may be provided by cloud computing, and cloud computing resources may be provided by one or more service providers.

**[0263]** In some examples, the device can include a storage and/or memory device. The storage and/or memory device may be one or more physical apparatuses used to store data or programs on a temporary or permanent basis. In some examples, the device may be volatile memory and require power to maintain stored information. In some examples, the device may be non-volatile memory and retain stored information when the digital processing device is not powered. In some examples, the non-volatile memory may include flash memory. In some examples, the non-volatile memory may include dynamic random-access memory (DRAM). In some examples, the non-volatile memory may include ferroelectric random access memory (FRAM). In some examples, the non-volatile memory may include phase-change random access memory (PRAM).

**[0264]** In some examples, the device may be a storage device including, for example, CD-ROMs, DVDs, flash memory devices, magnetic disk drives, magnetic tapes drives, optical disk drives, and cloud computing-based storage. In some examples, the storage and/or memory device may be a combination of devices such as those disclosed herein. In some examples, the

digital processing device may include a display to send visual information to a user. In some examples, the display may be a cathode ray tube (CRT). In some examples, the display may be a liquid crystal display (LCD). In some examples, the display may be a thin film transistor liquid crystal display (TFT-LCD). In some examples, the display may be an organic light emitting diode (OLED) display. In some examples, on OLED display may be a passive-matrix OLED (PMOLED) or active-matrix OLED (AMOLED) display. In some examples, the display may be a plasma display. In some examples, the display may be a video projector. In some examples, the display may be a combination of devices such as those disclosed herein.

**[0265]** In some examples, the digital processing device may include an input device to receive information from a user. In some examples, the input device may be a keyboard. In some examples, the input device may be a pointing device including, for example, a mouse, trackball, track pad, joystick, game controller, or stylus. In some examples, the input device may be a touch screen or a multi-touch screen. In some examples, the input device may be a microphone to capture voice or other sound input. In some examples, the input device may be a video camera to capture motion or visual input. In some examples, the input device may be a combination of devices such as those disclosed herein.

### D. Non-transitory computer-readable storage medium

**[0266]** In some examples, the subject matter disclosed herein may include one or more non-transitory computer-readable storage media encoded with a program including instructions executable by the operating system of an optionally networked digital processing device. In some examples, a computer-readable storage medium may be a tangible component of a digital processing device. In some examples, a computer-readable storage medium may be optionally removable from a digital processing device. In some examples, a computer-readable storage medium may include, for example, CD-ROMs, DVDs, flash memory devices, solid state memory, magnetic disk drives, magnetic tape drives, optical disk drives, cloud computing systems and services, and the like. In some examples, the program and instructions may be permanently, substantially permanently, semi-permanently, or non-transitorily encoded on the media.

### E. Computer systems

**[0267]** The present disclosure provides computer systems that are programmed to implement methods described herein. **FIG. 1** shows a computer system **101** that may be programmed or otherwise configured to store, process, identify, or interpret patient data, biological data, biological sequences, and reference sequences. The computer system **101** may process various aspects of patient data, biological data, biological sequences, or reference sequences of the

present disclosure (**FIG. 1**). The computer system **101** may be an electronic device of a user or a computer system that is remotely located with respect to the electronic device. The electronic device may be a mobile electronic device.

**[0268]** The computer system **101** may comprise a central processing unit (CPU, also "processor" and "computer processor" herein) **105**, which may be a single core or multi core processor, or a plurality of processors for parallel processing. The computer system **101** may also comprises memory or memory location **110** (e.g., random-access memory, read-only memory, flash memory), electronic storage unit **115** (e.g., hard disk), communication interface **120** (e.g., network adapter) for communicating with one or more other systems, and peripheral devices **125**, such as cache, other memory, data storage and/or electronic display adapters. The memory **110**, storage unit **115**, interface **120**, and peripheral devices **125** may be in communication with the CPU **105** through a communication bus (solid lines), such as a motherboard. The storage unit **115** may be a data storage unit (or data repository) for storing data. The computer system **101** may be operatively coupled to a computer network ("network") **130** with the aid of the communication interface **120**. The network **130** may be the Internet, an internet and/or extranet, or an intranet and/or extranet that is in communication with the Internet. The network **130**, in some examples, may be a telecommunication and/or data network. The network **130** may include one or more computer servers, which may enable distributed computing, such as cloud computing. The network **130**, in some examples with the aid of the computer system **101**, may implement a peer-to-peer network, which may enable devices coupled to the computer system **101** to behave as a client or a server.

**[0269]** The CPU **105** may execute a sequence of machine-readable instructions, which may be embodied in a program or software. The instructions may be stored in a memory location, such as the memory **110**. The instructions may be directed to the CPU **105**, which may subsequently program or otherwise configure the CPU **105** to implement methods of the present disclosure. Examples of operations performed by the CPU **105** may include fetch, decode, execute, and writeback.

**[0270]** The CPU **105** may be part of a circuit, such as an integrated circuit. One or more other components of the system **101** may be included in the circuit. In some examples, the circuit may be an application specific integrated circuit (ASIC).

**[0271]** The storage unit **115** may store files, such as drivers, libraries, and saved programs. The storage unit **115** may store user data, e.g., user preferences and user programs. The computer system **101**, in some examples, may include one or more additional data storage units that may be external to the computer system **101**, such as located on a remote server that is in communication with the computer system **101** through an intranet or the Internet.

**[0272]** The computer system **101** may communicate with one or more remote computer systems through the network **130**. For instance, the computer system **101** may communicate with a remote computer system of a user. Examples of remote computer systems may 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 may access the computer system **101** via the network **130**.

**[0273]** Methods as described herein may be implemented by way of machine (e.g., computer processor) executable code stored on an electronic storage location of the computer system **101**, such as, for example, on the memory **110** or electronic storage unit **115**. The machine-executable or machine-readable code may be provided in the form of software. During use, the code may be executed by the processor **105**. In some examples, the code may be retrieved from the storage unit **115** and stored on the memory **110** for ready access by the processor **105**. In some examples, the electronic storage unit **115** may be precluded, and machine-executable instructions are stored on memory **110**.

**[0274]** The code may be pre-compiled and configured for use with a machine having a processer adapted to execute the code or may be interpreted or compiled during runtime. The code may be supplied in a programming language that may be selected to enable the code to execute in a pre-compiled, interpreted, or as-compiled fashion.

**[0275]** Aspects of the systems and methods provided herein, such as the computer system **101**, may be embodied in programming. Various aspects of the technology may be thought of as "products" or "articles of manufacture" , for example, 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 may be stored on an electronic storage unit, such as memory (e.g., read-only memory, random-access memory, flash memory) or a hard disk. "Storage" type media may include any or all of the tangible memory of the computers, processors or the like, or associated modules thereof, such as various semiconductor memories, tape drives, disk drives and the like, which may provide non-transitory storage at any time for the software programming. All or portions of the software may at times be communicated through the Internet or various other telecommunication networks. Such communications, for example, may enable loading of the software from one computer or processor into another, for example, from a management server or host computer into the computer platform of an application server. Thus, another type of media that may bear the software elements comprises optical, electrical, and electromagnetic waves, such as used across physical interfaces between local devices, through wired and optical landline networks and over various air-links. The physical elements that carry such waves, such as wired or wireless links, optical links or the like,

also may be considered as media bearing the software. As used herein, unless restricted to non-transitory, tangible "storage" media, terms such as computer or machine "readable medium" may refer to any medium that participates in providing instructions to a processor for execution.

**[0276]** 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 may 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.

**[0277]** The computer system **101** may include or be in communication with an electronic display **135** that comprises a user interface (UI) **140** for providing, for example, a nucleic acid sequence, an enriched nucleic acid sample, a methylation profile, an expression profile, and an analysis of a methylation or expression profile. Examples of UI's may include, without limitation, a graphical user interface (GUI) and web-based user interface.

**[0278]** Methods and systems of the present disclosure may be implemented by way of one or more algorithms. An algorithm may be implemented by way of software upon execution by the central processing unit **105**. The algorithm can, for example, store, process, identify, or interpret patient data, biological data, biological sequences, and reference sequences.

**[0279]** While certain examples of methods and systems have been shown and described herein, one of skill in the art will realize that these are provided by way of example only and not intended to be limiting within the specification. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the scope described herein. Furthermore, it shall be understood that all aspects of the described methods and systems are not

limited to the specific depictions, configurations or relative proportions set forth herein which depend upon a variety of conditions and variables and the description is intended to include such alternatives, modifications, variations, or equivalents.

**[0280]** In some examples, the subject matter disclosed herein can include at least one computer program or use of the same. A computer program can a sequence of instructions, executable in the digital processing device's CPU, GPU, or TPU, written to perform a specified task. Computer-readable instructions may be implemented as program modules, such as functions, objects, Application Programming Interfaces (APIs), data structures, and the like, that perform particular tasks or implement particular abstract data types. In light of the disclosure provided herein, a computer program may be written in various versions of various languages.

**[0281]** The functionality of the computer-readable instructions may be combined or distributed as desired in various environments. In some examples, a computer program can include one sequence of instructions. In some examples, a computer program can include a plurality of sequences of instructions. In some examples, a computer program may be provided from one location. In some examples, a computer program may be provided from a plurality of locations. In some examples, a computer program can include one or more software modules. In some examples, a computer program can include, in part or in whole, one or more web applications, one or more mobile applications, one or more standalone applications, one or more web browser plug-ins, extensions, add-ins, or add-ons, or combinations thereof.

**[0282]** In some examples, the computer processing may be a method of statistics, mathematics, biology, or any combination thereof. In some examples, the computer processing method comprises a dimension reduction method including, for example, logistic regression, dimension reduction, principal component analysis, autoencoders, singular value decomposition, Fourier bases, singular value decomposition, wavelets, discriminant analysis, support vector machine, tree-based methods, random forest, gradient boost tree, logistic regression, matrix factorization, network clustering, and neural network such as convolutional neural networks.

**[0283]** In some examples, the computer processing method may be a supervised machine learning method including, for example, a regression, support vector machine, tree-based method, and network.

**[0001]** In some examples, the computer processing method may be an unsupervised machine learning method including, for example, clustering, network, principal component analysis, and matrix factorization.

**F. Databases**

**[0284]** In some examples, the subject matter disclosed herein may include one or more databases, or use of the same to store patient data, biological data, biological sequences, or reference sequences. Reference sequences may be derived from a database. In view of the disclosure provided herein, many databases may be suitable for storage and retrieval of the sequence information. In some examples, suitable databases can include, for example, relational databases, non-relational databases, object-oriented databases, object databases, entity-relationship model databases, associative databases, and XML databases. In some examples, a database may be internet-based. In some examples, a database may be web-based. In some examples, a database may be cloud computing-based. In some examples, a database may be based on one or more local computer storage devices.

**[0285]** In an aspect, the present disclosure provides a non-transitory computer-readable medium comprising instructions that direct a processor to carry out a method disclosed herein.

**[0286]** In an aspect, the present disclosure provides a computing device comprising the computer-readable medium.

**[0287]** In another aspect, the present disclosure provides a system for performing classifications of biological samples comprising:

a) a receiver to receive a plurality of training samples, each of the plurality of training samples having a plurality of classes of molecules, wherein each of the plurality of training samples comprises one or more known labels;

b) a feature module to identify a set of features corresponding to an assay that are operable to be analyzed using the machine learning model for each of the plurality of training samples, wherein the set of features corresponds to properties of molecules in the plurality of training samples, wherein for each of the plurality of training samples, the system is operable to subject a plurality of classes of molecules in the training samples to a plurality of different assays to obtain sets of measured values, wherein each set of measured values is from one assay applied to a class of molecules in the training samples, wherein a plurality of sets of measured values are obtained for the plurality of training samples;

c) an analysis module to analyze the sets of measured values to obtain a training vector for the training samples, wherein the training vector comprises feature values of an $N$ set of features of the corresponding assay, each feature value corresponding to a feature and including one or more measured values, wherein the training vector is formed using at least one feature from at least two of the $N$ sets of features corresponding to a first subset of the plurality of different assays;

d) a labeling module to inform the system on the training vectors using parameters of the machine learning model to obtain output labels for the plurality of training samples;

e) a comparator module to compare the output labels to the known labels of the training samples;

f) a training module to iteratively search for optimal values of the parameters as part of training the machine learning model based on the comparing the output labels to the known labels of the training samples; and

g) an output module to provide the parameters of the machine learning model and the set of features for the machine learning model.

## VI. METHODS OF CLASSIFYING SUBJECTS IN A POPULATION

**[0288]** The disclosed methods are directed to ascertaining genetic and/or epigenetic parameters of genomic DNA associated with cell proliferative disorders via analysis of cfDNA in a subject. The method may be for use in the improved diagnosis, treatment, and monitoring of cell proliferative disorders, more specifically by enabling the improved identification of and differentiation between stages or subclasses of said disorder and the genetic predisposition to said disorders.

**[0289]** In some embodiments, the method comprises analyzing the methylation status of CpG islands, CpG shores, or CpG shelves.

**[0290]** In some embodiments, the method comprises analyzing the methylation state, hemimethylation status, hypermethylation state, or hypomethylation state of a cell-free nucleic acid in a biological sample.

**[0291]** Generally, the present disclosure provides a method for detecting a cell proliferative disorder that may be applied to cell-free samples, e.g., to detect cell-free circulating cell proliferative disorder DNA. The method may utilize detection of methylation signals within a single sequencing read as the basic “positive” cell proliferative disorder signal.

**[0292]** In an aspect, the present disclosure provides a method for detecting a cell proliferative disorder, comprising: extracting DNA from a cell-free sample obtained from a subject, converting at least a portion of the DNA for methyl sequencing, amplifying regions methylated in cancer from the converted DNA, generating sequencing reads from the amplified regions, and detecting cell proliferative disorder signals comprising at least one, at least two, at least three, or more than three methylated regions within a cancer panel, to obtain input features that may be analyzed using a machine learning model to obtain a classifier capable of discriminating between two groups of subjects (e.g., healthy vs. cancer, disease stage, advanced adenoma vs. cancer).

**[0293]** The trained machine learning methods, models, and discriminate classifiers described herein may be applied toward various medical applications including cancer detection,

diagnosis, and treatment responsiveness. As models may be trained with individual metadata and analyte-derived features, the applications may be tailored to stratify individuals in a population and guide treatment decisions accordingly.

***Diagnosis***

**[0294]** Methods and systems provided herein may perform predictive analytics using artificial intelligence-based approaches to analyze acquired data from a subject (or patient) to generate an output of diagnosis of the subject having a cancer. For example, the application may apply a prediction algorithm to the acquired data to generate the diagnosis of the subject having the cancer. The prediction algorithm may comprise an artificial intelligence-based predictor, such as a machine learning-based predictor, configured to process the acquired data to generate the diagnosis of the subject having the cancer.

**[0295]** The machine learning predictor may be trained using datasets, e.g., datasets generated by performing methylation assays using the signature panels described herein on biological samples of individuals from one or more sets of cohorts of patients having cancer as inputs and known diagnosis (e.g., staging and/or tumor fraction) outcomes of the subjects as outputs to the machine learning predictor.

**[0296]** Training datasets (e.g., datasets generated by performing methylation assays using the signature panels described herein on biological samples of individuals) may be generated from, for example, one or more sets of subjects having common characteristics (features) and outcomes (labels). Training datasets may comprise a set of features and labels corresponding to the features relating to diagnosis. Features may comprise characteristics such as, for example, certain ranges or categories of cfDNA assay measurements, such as counts of cfDNA fragments in a biological sample obtained from a healthy and disease samples that overlap or fall within each of a set of bins (genomic windows) of a reference genome. For example, a set of features collected from a given subject at a given time point may collectively serve as a diagnostic signature, which may be indicative of an identified cancer of the subject at the given time point. Characteristics may also include labels indicating the subject's diagnostic outcome, such as for one or more cancers.

**[0297]** Labels may comprise outcomes such as, for example, a known diagnosis (e.g., staging and/or tumor fraction) outcomes of the subject. Outcomes may include a characteristic associated with the cancers in the subject. For example, characteristics may be indicative of the subject having one or more cancers.

**[0298]** Training sets (e.g., training datasets) may be selected by random sampling of a set of data corresponding to one or more sets of subjects (e.g., retrospective and/or prospective cohorts

of patients having or not having one or more cancers). Alternatively, training sets (e.g., training datasets) may be selected by proportionate sampling of a set of data corresponding to one or more sets of subjects (e.g., retrospective and/or prospective cohorts of patients having or not having one or more cancers). Training sets may be balanced across sets of data corresponding to one or more sets of subjects (e.g., patients from different clinical sites or trials). The machine learning predictor may be trained until certain predetermined conditions for accuracy or performance are satisfied, such as having minimum desired values corresponding to diagnostic accuracy measures. For example, the diagnostic accuracy measure may correspond to prediction of a diagnosis, staging, or tumor fraction of one or more cancers in the subject.

**[0299]** Examples of diagnostic accuracy measures may include sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and area under the curve (AUC) of a Receiver Operating Characteristic (ROC) curve corresponding to the diagnostic accuracy of detecting or predicting the cancer.

**[0300]** In an aspect, the disclosure provides a method of using a classifier capable of distinguishing a population of individuals, comprising:

a) assaying a plurality of classes of molecules in the biological sample, wherein the assaying provides a plurality of sets of measured values representative of the plurality of classes of molecules;

b) identifying a set of features corresponding to properties of each of the plurality of classes of molecules to be analyzed using a machine learning or statistical model;

c) preparing a feature vector of feature values from each of the plurality of sets of measured values, each feature value corresponding to a feature of the set of features and including one or more measured values, wherein the feature vector comprises at least one feature value obtained using each set of the plurality of sets of measured values;

d) loading, into a memory of a computer system, the machine learning model comprising the classifier, the machine learning model trained using training vectors obtained from training biological samples, a first subset of the training biological samples identified as having a specified property and a second subset of the training biological samples identified as not having the specified property; and

e) analyzing the feature vector using the machine learning model to obtain an output classification of whether the biological sample has the specified property, thereby distinguishing a population of individuals having the specified property.

**[0301]** In an aspect, the disclosure provides a method of using a hierarchy capable of distinguishing a population of individuals comprising:

a) assaying a plurality of classes of molecules in the biological sample, wherein the assaying provides a plurality of sets of measured values representative of the plurality of classes of molecules;

b) identifying a set of features corresponding to properties of each of the plurality of classes of molecules to be analyzed using a machine learning or statistical model;

c) preparing a feature vector of feature values from each of the plurality of sets of measured values, each feature value corresponding to a feature of the set of features and including one or more measured values, wherein the feature vector comprises at least one feature value obtained using each set of the plurality of sets of measured values;

d) loading, into a memory of a computer system, a trained machine learning model comprising the classifier, the trained machine learning model trained using training vectors obtained from training biological samples, a first subset of the training biological samples identified as having a specified property and a second subset of the training biological samples identified as not having the specified property; and

e) applying the trained machine learning model to the feature vector to obtain an output classification of whether the biological sample has the specified property, thereby distinguishing a population of individuals having the specified property.

**[0302]** In an aspect, the disclosure provides a method of using a hierarchy capable of distinguishing a population of individuals, comprising:

a) detecting of methylation signals within a single sequencing read of a pre-selected genomic region in one or more first patient samples;

b) the methylation signals affect a hierarchy of data outputs to affect a machine learning model; and

c) a second patient sample using the affected hierarchy to detect methylation signals.

**[0303]** In some embodiments, the signature panel comprises three or more methylated genomic regions in **Tables 2-17**, four or more methylated genomic regions in **Tables 2-17**, five or more methylated genomic regions in **Tables 2-17**, six or more methylated genomic regions in **Tables 2-17**, seven or more methylated genomic regions in **Tables 2-17**, eight or more methylated genomic regions in **Tables 2-17**, nine or more methylated genomic regions in **Tables 2-17**, ten or more methylated genomic regions in **Tables 2-17**, eleven or more methylated genomic regions in **Tables 2-17**, twelve or more methylated genomic regions in **Tables 2-17**, or thirteen or more methylated genomic regions in **Tables 2-17**.

**[0304]** In another aspect, the present disclosure provides a method for identifying two or more cancers in a subject, comprising:

(a) providing a biological sample comprising cell-free nucleic acid (cfNA) molecules from said subject;

(b) methyl converting and sequencing said cfNA molecules from said subject to generate a plurality of cfNA sequencing reads;

(c) aligning said plurality of cfNA sequencing reads to a reference genome;

(d) generating a quantitative measure of said plurality of cfNA sequencing reads at each of a first plurality of genomic regions of said reference genome to generate a first cfNA feature set, wherein said first plurality of genomic regions of said reference genome comprises at least about 10 distinct regions, each of said at least about 10 distinct regions comprising at least a portion of a gene selected from the group consisting of methylated regions in the signature panels described herein; and

(e) applying a trained algorithm to said first cfNA feature set to generate a likelihood of said subject having said cancer.

**[0305]** In some examples, said at least about 10 distinct regions comprises at least about 20 distinct regions, each of said at least about 20 distinct regions comprising at least a portion of a methylated region identified in **Tables 1-17**. In some examples, said at least about 10 distinct regions comprises at least about 30 distinct regions, each of said at least about 30 distinct regions comprising at least a portion of a methylated region identified in **Tables 1-17**.

**[0306]** As another example, such a predetermined condition may be that the specificity of predicting the colon cell proliferative disorder comprises a value of, for example, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, or at least about 99%.

**[0307]** As another example, such a predetermined condition may be that the positive predictive value (PPV) of predicting the colon cell proliferative disorder comprises a value of, for example, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, or at least about 99%.

**[0308]** As another example, such a predetermined condition may be that the negative predictive value (NPV) of predicting the colon cell proliferative disorder comprises a value of, for example, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, or at least about 99%.

**[0309]** As another example, such a predetermined condition may be that the area under the curve (AUC) of a Receiver Operating Characteristic (ROC) curve of predicting the colon cell proliferative disorder comprises a value of at least about 0.50, at least about 0.55, at least about 0.60, at least about 0.65, at least about 0.70, at least about 0.75, at least about 0.80, at least about 0.85, at least about 0.90, at least about 0.95, at least about 0.96, at least about 0.97, at least about 0.98, or at least about 0.99.

***Treatment Responsiveness***

**[0310]** The predictive classifiers, systems, and methods described herein may be applied toward classifying populations of individuals for a number of clinical applications (e.g., based on performing methylation assays using the signature panels described herein on biological samples of individuals). Examples of such clinical applications include detecting early-stage cancer, diagnosing cancer, classifying cancer to a particular stage of disease, and determining responsiveness or resistance to a therapeutic agent for treating cancer.

**[0311]** The methods and systems described herein may be applied to characteristics of a colon cell proliferative disorder, such as grade and stage. Therefore, combinations of analytes and assays may be used in the present systems and methods to predict responsiveness of cancer therapeutics across different cancer types in different tissues and classifying individuals based on treatment responsiveness. In some embodiments, the classifiers described herein are capable of stratifying a group of individuals into treatment responders and non-responders.

**[0312]** The present disclosure also provides a method for determining a drug target of a condition or disease of interest (e.g., genes that are relevant or important for a particular class), comprising: assessing a sample obtained from an individual for the level of gene expression for at least one gene; and using a neighborhood analysis routine, determining genes that are relevant for classification of the sample, to thereby ascertain one or more drug targets relevant to the classification.

**[0313]** The present disclosure also provides a method for determining the efficacy of a drug designed to treat a disease class, comprising obtaining a sample from an individual having the disease class; subjecting the sample to the drug; assessing the drug-exposed sample for the level of gene expression for at least one gene; and using a computer model built with a weighted voting scheme, classifying the drug-exposed sample into a class of the disease as a function of relative gene expression level of the sample with respect to that of the model.

**[0314]** The present disclosure also provides a method for determining the efficacy of a drug designed to treat a disease class, wherein an individual has been subjected to the drug, comprising obtaining a sample from the individual subjected to the drug; assessing the sample

for the level of gene expression for at least one gene; and using a model built with a weighted voting scheme, classifying the sample into a class of the disease including evaluating the gene expression level of the sample as compared to gene expression level of the model.

**[0315]** The present disclosure also provides a method of determining whether an individual belongs to a phenotypic class (e.g., intelligence, response to a treatment, length of life, likelihood of viral infection, or obesity), comprising obtaining a sample from the individual; assessing the sample for the level of gene expression for at least one gene; and using a model built with a weighted voting scheme, classifying the sample into a class of the disease including evaluating the gene expression level of the sample as compared to gene expression level of the model.

**[0316]** In an aspect, the systems and methods described herein that relate to classifying a population based on treatment responsiveness refer to cancers that are treated with chemotherapeutic agents of the classes DNA damaging agents, DNA repair target therapies, inhibitors of DNA damage signaling, inhibitors of DNA damage induced cell cycle arrest and inhibition of processes indirectly leading to DNA damage, but not limited to these classes. Each of these chemotherapeutic agents may be considered a "DNA-damage therapeutic agent" as the term is used herein.

**[0317]** Based on a patient's analyte data, the patient may be classified into high-risk and low-risk patient groups, such as patient with a high or low risk of clinical relapse, and the results may be used to determine a course of treatment. For example, a patient determined to be a high-risk patient may be treated with adjuvant chemotherapy after surgery. For a patient deemed to be a low-risk patient, adjuvant chemotherapy may be withheld after surgery. Accordingly, the present disclosure provides, in certain aspects, a method for preparing a gene expression profile of a colon cancer tumor that is indicative of risk of recurrence.

**[0318]** In various examples, the classifiers described herein are capable of stratifying a population of individuals between responders and non-responders to treatment.

**[0319]** In another aspect, methods disclosed herein may be applied to clinical applications involving the detection or monitoring of cancer.

**[0320]** In some embodiments, methods disclosed herein may be applied to determine and/or predict response to treatment.

**[0321]** In some embodiments, methods disclosed herein may be applied to monitor and/or predict tumor load.

**[0322]** In some embodiments, methods disclosed herein may be applied to detect and /or predict residual tumor post-surgery.

**[0323]** In some embodiments, methods disclosed herein may be applied to detect and /or predict minimal residual disease post-treatment.

**[0324]** In some embodiments, methods disclosed herein may be applied to detect and/or predict relapse.

**[0325]** In an aspect, methods disclosed herein may be applied as a secondary screen.

**[0326]** In an aspect, methods disclosed herein may be applied as a primary screen.

**[0327]** In an aspect, methods disclosed herein may be applied to monitor cancer development.

**[0328]** In an aspect, methods disclosed herein may be applied to monitor and/or predict cancer risk.

## VII. IDENTIFYING OR MONITORING CANCER

**[0329]** After using a trained algorithm to process the dataset, at least two cancer types may be identified or monitored in the subject. The identification may be based at least in part on quantitative measures of sequence reads of the dataset at a panel of cancer-associated genomic loci (e.g., quantitative measures of RNA transcripts or DNA at the cancer-associated genomic loci).

**[0330]** In one embodiment, 2 or more cancer types are identified or monitored in the subject, in another embodiment, 3 or more cancer types are identified or monitored in the subject, in another embodiment, 4 or more cancer types are identified or monitored in the subject, in another embodiment, 5 or more cancer types are identified or monitored in the subject, in another embodiment, 6 or more cancer types are identified or monitored in the subject, in another embodiment, 7 or more cancer types are identified or monitored in the subject, in another embodiment, 8 or more cancer types are identified or monitored in the subject, in another embodiment, 9 or more cancer types are identified or monitored in the subject, in another embodiment, 10 or more cancer types are identified or monitored in the subject.

**[0331]** The cancer may be identified in the subject at an accuracy of at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 81%, at least about 82%, at least about 83%, at least about 84%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or more. The accuracy of identifying the cancer by the trained algorithm may be calculated as the percentage of independent test samples (e.g., subjects known to have the cancer or subjects with negative clinical test results for the cancer) that are correctly identified or classified as having or not having the cancer.

**[0332]** The cancer may be identified in the subject with a positive predictive value (PPV) of at least about 5%, at least about 10%, at least about 15%, at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 81%, at least about 82%, at least about 83%, at least about 84%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or more. The PPV of identifying the cancer using the trained algorithm may be calculated as the percentage of cell-free biological samples identified or classified as having the cancer that correspond to subjects that truly have the cancer.

**[0333]** The cancer may be identified in the subject with a negative predictive value (NPV) of at least about 5%, at least about 10%, at least about 15%, at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 81%, at least about 82%, at least about 83%, at least about 84%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or more. The NPV of identifying the cancer using the trained algorithm may be calculated as the percentage of cell-free biological samples identified or classified as not having the cancer that correspond to subjects that truly do not have the cancer.

**[0334]** The cancer may be identified in the subject with a clinical sensitivity of at least about 5%, at least about 10%, at least about 15%, at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 81%, at least about 82%, at least about 83%, at least about 84%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, at least about 99.1%, at least about 99.2%, at least about 99.3%, at least about 99.4%, at least about 99.5%, at least about 99.6%, at least about 99.7%, at least about 99.8%, at least about 99.9%, at least about 99.99%, at least about 99.999%, or more. The clinical sensitivity of identifying the cancer using the trained algorithm may be calculated as the percentage of independent test

samples associated with presence of the cancer (e.g., subjects known to have the cancer) that are correctly identified or classified as having the cancer.

**[0335]** The cancer may be identified in the subject with a clinical specificity of at least about 5%, at least about 10%, at least about 15%, at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 81%, at least about 82%, at least about 83%, at least about 84%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, at least about 99.1%, at least about 99.2%, at least about 99.3%, at least about 99.4%, at least about 99.5%, at least about 99.6%, at least about 99.7%, at least about 99.8%, at least about 99.9%, at least about 99.99%, at least about 99.999%, or more. The clinical specificity of identifying the cancer using the trained algorithm may be calculated as the percentage of independent test samples associated with absence of the cancer (e.g., subjects with negative clinical test results for the cancer) that are correctly identified or classified as not having the cancer.

**[0336]** In some embodiments, the trained algorithm may determine that the subject is at risk of cancer of at least about 5%, at least about 10%, at least about 15%, at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 81%, at least about 82%, at least about 83%, at least about 84%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or more.

**[0337]** The trained algorithm may determine that the subject is at risk of cancer at an accuracy of at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 81%, at least about 82%, at least about 83%, at least about 84%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, at least about 99.1%, at least about 99.2%, at least about 99.3%, at least about 99.4%, at least about 99.5%, at least about 99.6%, at least about 99.7%, at least about 99.8%, at least about 99.9%, at least about 99.99%, at least about 99.999%, or more.

### A. Tailored Multicancer Signature Panels

**[0338]** In some embodiments, a multicancer detection assay biomarker panel comprises test characteristics that are selected for the different cancer types assayed in the signature panel and in subsequent analysis. In certain embodiments, the test characteristics may be ascertained from screening goals and signature panel marker selection. For example, for a first line screening test, some cancers may require greater sensitivity at a clinically acceptable specificity, while others may require very high specificity at a clinically acceptable sensitivity due to the benefits and risks of the subsequent diagnostic workup. Furthermore, performance characteristics depend on whether the test precedes, complements, or follows an accepted method of screening, or represents a new frontline screen for an otherwise unscreened cancer, either in an asymptomatic, average-risk or symptomatic, high-risk individual. For example, the impact to the patient of a false positive screen for colorectal cancer (CRC) resulting in an "unnecessary" colonoscopy is meaningfully different from a false positive screen for pancreatic or ovarian cancer that results in the "unnecessary" major abdominal surgery to confirm diagnosis. When combined with signature panel marker selection, multicancer detection biomarker panels provide methods and systems that are tailored for the screening goals, confirmatory tests, and subsequent treatment available.

**[0339]** **Table 18** summarizes screening test characteristics for multiple cancer detection tests. In an aspect, a method is provided where the multicancer panel is tailored to provide test characteristic sensitivity and specificity for the types of cancer to be detected based on needs of cancer diagnosis and confirmatory diagnosis for two or more of the cancer types shown in **Table 18** or combinations thereof.

**Table 18**

| Cancer Type | Screening Goal | Conventional Diagnostic | Multicancer Test Characteristic |
|---|---|---|---|
| CRC | Minimize false negatives; avoid unnecessary screening colonoscopies | Colonoscopy | High NPV (high sensitivity) |
| Breast | Minimize false positives; avoid unnecessary biopsies and mastectomies | Fine needle aspirate or core biopsy | High PPV (high specificity) |
| Ovarian | Minimize false positives; avoid unnecessary major abdominal surgery | Abdominal Surgery | Very High PPV (very high specificity) |
| Prostate | Minimize false positives; avoid unnecessary biopsies | Biopsy | High PPV (high specificity) |
| Lung | Minimize false positives; avoid unnecessary and expensive imaging | Imaging (X-ray, CT scan); sputum cytology; tissue biopsy | High PPV (high specificity) |

| | | | |
|---|---|---|---|
| Pancreatic | Minimize false negatives | Abdominal Surgery | Very High PPV (very high specificity) |
| Uterine | Minimize false positives; avoid unnecessary major abdominal surgery | Abdominal Surgery | Very High PPV (very high specificity) |
| Liver | Minimize false negatives | Imaging and Biopsy | High NPV (high sensitivity) |
| Esophagus | Minimize false negatives; accurately stage cancer to select appropriate treatment | Biopsy | High NPV (high sensitivity) |
| Stomach | Minimize false negatives | Endoscopic biopsy | High NPV (high sensitivity) |
| Thyroid | Minimize false positives; avoid unnecessary biopsies | fine-needle aspiration | High PPV (high specificity) |
| Bladder | Minimize false negatives | cystoscopy | High NPV (high sensitivity) |

**[0340]** In one embodiment, the multicancer test comprises markers for detecting pancreatic, uterine, or ovarian cancer, and has a specificity at least 80%, at least 85%, at least 90%, at least 95%, at least 99%.

**[0341]** In one embodiment, the multicancer test comprises markers for detecting colorectal, liver, esophagus, or bladder cancer, and has a sensitivity of at least 50%, at least 60%, at least 70%, at least 80%, at least 90%, at least 95%.

**[0342]** In one embodiment, the multicancer test comprises markers for detecting breast, prostate, lung, or thyroid cancer, and has a specificity of at least 50%, at least 60%, at least 70%, at least 80%, at least 90%, at least 95%.

**[0343]** Upon identifying the subject as having a cancer type, the subject may be optionally provided with a therapeutic intervention (e.g., prescribing an appropriate course of treatment to treat the cancer of the subject). The therapeutic intervention may comprise a prescription of an effective dose of a drug, a further testing or evaluation of the cancer, a further monitoring of the cancer, or a combination thereof. If the subject is currently being treated for the cancer with a course of treatment, the therapeutic intervention may comprise a subsequent different course of treatment (e.g., to increase treatment efficacy due to non-efficacy of the current course of treatment).

**[0344]** The therapeutic intervention may comprise recommending the subject for a secondary clinical test to confirm a diagnosis of the cancer. This secondary clinical test may comprise an imaging test, a blood test, a computed tomography (CT) scan, a magnetic resonance imaging (MRI) scan, an ultrasound scan, a chest X-ray, a positron emission tomography (PET) scan, a

PET-CT scan, a cell-free biological cytology, a FIT test, an FOBT test, or any combination thereof.

**[0345]** The quantitative measures of sequence reads of the dataset at the panel of cancer-associated genomic loci (e.g., quantitative measures of RNA transcripts or DNA at the colorectal cancer-associated genomic loci) may be assessed over a duration of time to monitor a patient (e.g., subject who has cancer or who is being treated for cancer). In such cases, the quantitative measures of the dataset of the patient may change during the course of treatment. For example, the quantitative measures of the dataset of a patient with decreasing risk of the cancer due to an effective treatment may shift toward the profile or distribution of a healthy subject (e.g., a subject without cancer). Conversely, for example, the quantitative measures of the dataset of a patient with increasing risk of the cancer due to an ineffective treatment may shift toward the profile or distribution of a subject with higher risk of the cancer or a more advanced cancer.

**[0346]** The cancer of the subject may be monitored by monitoring a course of treatment for treating the cancer of the subject. The monitoring may comprise assessing the cancer of the subject at two or more time points. The assessing may be based at least on the quantitative measures of sequence reads of the dataset at a panel of cancer-associated genomic loci (e.g., quantitative measures of RNA transcripts or DNA at the cancer-associated genomic loci) comprising quantitative measures of a panel of cancer-associated genomic loci determined at each of the two or more time points.

**[0347]** In some embodiments, a difference in the quantitative measures of sequence reads of the dataset at a panel of cancer-associated genomic loci (e.g., quantitative measures of RNA transcripts or DNA at the cancer-associated genomic loci) comprising quantitative measures of a panel of cancer-associated genomic loci determined between the two or more time points may be indicative of one or more clinical indications, such as (i) a diagnosis of the cancer of the subject; (ii) a prognosis of the cancer of the subject; (iii) an increased risk of the cancer of the subject; (iv) a decreased risk of the cancer of the subject; (v) an efficacy of the course of treatment for treating the cancer of the subject; and (vi) a non-efficacy of the course of treatment for treating the cancer of the subject.

**[0348]** In some embodiments, a difference in the quantitative measures of sequence reads of the dataset at a panel of cancer-associated genomic loci (e.g., quantitative measures of RNA transcripts or DNA at the cancer-associated genomic loci) comprising quantitative measures of a panel of cancer-associated genomic loci determined between the two or more time points may be indicative of a diagnosis of the cancer of the subject. For example, if the cancer was not detected in the subject at an earlier time point but was detected in the subject at a later time point, then the difference is indicative of a diagnosis of the cancer of the subject. A clinical action or

decision may be made based on this indication of diagnosis of the cancer of the subject, such as, for example, prescribing a new therapeutic intervention for the subject. The clinical action or decision may comprise recommending the subject for a secondary clinical test to confirm the diagnosis of the cancer. This secondary clinical test may comprise an imaging test, a blood test, a computed tomography (CT) scan, a magnetic resonance imaging (MRI) scan, an ultrasound scan, a chest X-ray, a positron emission tomography (PET) scan, a PET-CT scan, a cell-free biological cytology, a FIT test, an FOBT test, or any combination thereof.

**[0349]** In some embodiments, a difference in the quantitative measures of sequence reads of the dataset at a panel of cancer-associated genomic loci (e.g., quantitative measures of RNA transcripts or DNA at the cancer-associated genomic loci) comprising quantitative measures of a panel of cancer-associated genomic loci determined between the two or more time points may be indicative of a prognosis of the cancer of the subject.

**[0350]** In some embodiments, a difference in the quantitative measures of sequence reads of the dataset at a panel of cancer-associated genomic loci (e.g., quantitative measures of RNA transcripts or DNA at the cancer-associated genomic loci) comprising quantitative measures of a panel of cancer-associated genomic loci determined between the two or more time points may be indicative of the subject having an increased risk of the cancer. For example, if the colorectal cancer was detected in the subject both at an earlier time point and at a later time point, and if the difference is a positive difference (e.g., the quantitative measures of sequence reads of the dataset at a panel of cancer-associated genomic loci (e.g., quantitative measures of RNA transcripts or DNA at the cancer-associated genomic loci) increased from the earlier time point to the later time point), then the difference may be indicative of the subject having an increased risk of the cancer. A clinical action or decision may be made based on this indication of the increased risk of the cancer, e.g., prescribing a new therapeutic intervention or switching therapeutic interventions (e.g., ending a current treatment and prescribing a new treatment) for the subject. The clinical action or decision may comprise recommending the subject for a secondary clinical test to confirm the increased risk of the cancer. This secondary clinical test may comprise an imaging test, a blood test, a computed tomography (CT) scan, a magnetic resonance imaging (MRI) scan, an ultrasound scan, a chest X-ray, a positron emission tomography (PET) scan, a PET-CT scan, a cell-free biological cytology, a FIT test, an FOBT test, or any combination thereof.

**[0351]** In some embodiments, a difference in the quantitative measures of sequence reads of the dataset at a panel of cancer-associated genomic loci (e.g., quantitative measures of RNA transcripts or DNA at the colorectal cancer-associated genomic loci) comprising quantitative measures of a panel of cancer-associated genomic loci determined between the two or more time

points may be indicative of the subject having a decreased risk of the cancer. For example, if the cancer was detected in the subject both at an earlier time point and at a later time point, and if the difference is a negative difference (e.g., the quantitative measures of sequence reads of the dataset at a panel of cancer-associated genomic loci (e.g., quantitative measures of RNA transcripts or DNA at the colorectal cancer-associated genomic loci) comprising quantitative measures of a panel of cancer-associated genomic loci decreased from the earlier time point to the later time point), then the difference may be indicative of the subject having a decreased risk of the colorectal cancer. A clinical action or decision may be made based on this indication of the decreased risk of the cancer (e.g., continuing or ending a current therapeutic intervention) for the subject. The clinical action or decision may comprise recommending the subject for a secondary clinical test to confirm the decreased risk of the colorectal cancer. This secondary clinical test may comprise an imaging test, a blood test, a computed tomography (CT) scan, a magnetic resonance imaging (MRI) scan, an ultrasound scan, a chest X-ray, a positron emission tomography (PET) scan, a PET-CT scan, a cell-free biological cytology, a FIT test, an FOBT test, or any combination thereof.

**[0352]** In some embodiments, a difference in the quantitative measures of sequence reads of the dataset at a panel of cancer-associated genomic loci (e.g., quantitative measures of RNA transcripts or DNA at the cancer-associated genomic loci) comprising quantitative measures of a panel of cancer-associated genomic loci determined between the two or more time points may be indicative of an efficacy of the course of treatment for treating the cancer of the subject. For example, if the cancer was detected in the subject at an earlier time point but was not detected in the subject at a later time point, then the difference may be indicative of an efficacy of the course of treatment for treating the cancer of the subject. A clinical action or decision may be made based on this indication of the efficacy of the course of treatment for treating the cancer of the subject, e.g., continuing or ending a current therapeutic intervention for the subject. The clinical action or decision may comprise recommending the subject for a secondary clinical test to confirm the efficacy of the course of treatment for treating the cancer. This secondary clinical test may comprise an imaging test, a blood test, a computed tomography (CT) scan, a magnetic resonance imaging (MRI) scan, an ultrasound scan, a chest X-ray, a positron emission tomography (PET) scan, a PET-CT scan, a cell-free biological cytology, a FIT test, an FOBT test, or any combination thereof.

**[0353]** In some embodiments, a difference in the quantitative measures of sequence reads of the dataset at a panel of cancer-associated genomic loci (e.g., quantitative measures of RNA transcripts or DNA at the cancer-associated genomic loci) comprising quantitative measures of a panel of cancer-associated genomic loci determined between the two or more time points may be

indicative of a non-efficacy of the course of treatment for treating the cancer of the subject. For example, if the cancer was detected in the subject both at an earlier time point and at a later time point, and if the difference is a positive or zero difference (e.g., the quantitative measures of sequence reads of the dataset at a panel of cancer-associated genomic loci (e.g., quantitative measures of RNA transcripts or DNA at the cancer-associated genomic loci) comprising quantitative measures of a panel of cancer-associated genomic loci increased or remained at a constant level from the earlier time point to the later time point), and if an efficacious treatment was indicated at an earlier time point, then the difference may be indicative of a non-efficacy of the course of treatment for treating the cancer of the subject. A clinical action or decision may be made based on this indication of the non-efficacy of the course of treatment for treating the cancer of the subject, e.g., ending a current therapeutic intervention and/or switching to (e.g., prescribing) a different new therapeutic intervention for the subject. The clinical action or decision may comprise recommending the subject for a secondary clinical test to confirm the non-efficacy of the course of treatment for treating the cancer. This secondary clinical test may comprise an imaging test, a blood test, a computed tomography (CT) scan, a magnetic resonance imaging (MRI) scan, an ultrasound scan, a chest X-ray, a positron emission tomography (PET) scan, a PET-CT scan, a cell-free biological cytology, a FIT test, an FOBT test, or any combination thereof.

## VIII. KITS

**[0354]** The present disclosure provides kits for identifying or monitoring two or more cancer types in a subject. A kit may comprise probes for identifying a quantitative measure (e.g., indicative of a presence, absence, or relative amount) of sequences at each of a plurality of cancer-associated genomic loci in a cell-free biological sample of the subject. A quantitative measure (e.g., indicative of a presence, absence, or relative amount) of sequences at each of a plurality of cancer-associated genomic loci in the cell-free biological sample may be indicative of one or more cancers. The probes may be selective for the sequences at the plurality of cancer-associated genomic loci in the cell-free biological sample. A kit may comprise instructions for using the probes to process the cell-free biological sample to generate datasets indicative of a quantitative measure (e.g., indicative of a presence, absence, or relative amount) of sequences at each of the plurality of cancer-associated genomic loci in a cell-free biological sample of the subject.

**[0355]** The probes in the kit may be selective for the sequences at the plurality of cancer-associated genomic loci in the cell-free biological sample. The probes in the kit may be configured to selectively enrich nucleic acid (e.g., RNA or DNA) molecules corresponding to

the plurality of cancer-associated genomic loci. The probes in the kit may be nucleic acid primers. The probes in the kit may have sequence complementarity with nucleic acid sequences from one or more of the plurality of cancer-associated genomic loci or genomic regions. The plurality of cancer-associated genomic loci or genomic regions may comprise at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, or more distinct cancer-associated genomic loci or genomic regions. The plurality of cancer-associated genomic loci or genomic regions may comprise one or more members selected from the group consisting of regions listed in **Tables 1-17**.

**[0356]** The instructions in the kit may comprise instructions to assay the cell-free biological sample using the probes that are selective for the sequences at the plurality of cancer-associated genomic loci in the cell-free biological sample. These probes may be nucleic acid molecules (e.g., RNA or DNA) having sequence complementarity with nucleic acid sequences (e.g., RNA or DNA) from one or more of the plurality of cancer-associated genomic loci. These nucleic acid molecules may be primers or enrichment sequences. The instructions to assay the cell-free biological sample may comprise introductions to perform array hybridization, polymerase chain reaction (PCR), or nucleic acid sequencing (e.g., DNA sequencing or RNA sequencing) to process the cell-free biological sample to generate datasets indicative of a quantitative measure (e.g., indicative of a presence, absence, or relative amount) of sequences at each of the plurality of cancer-associated genomic loci in the cell-free biological sample. A quantitative measure (e.g., indicative of a presence, absence, or relative amount) of sequences at each of a plurality of cancer-associated genomic loci in the cell-free biological sample may be indicative of one or more cancers.

**[0357]** The instructions in the kit may comprise instructions to measure and interpret assay readouts, which may be quantified at one or more of the plurality of cancer-associated genomic loci to generate the datasets indicative of a quantitative measure (e.g., indicative of a presence, absence, or relative amount) of sequences at each of the plurality of cancer-associated genomic loci in the cell-free biological sample. For example, quantification of array hybridization or polymerase chain reaction (PCR) corresponding to the plurality of cancer-associated genomic loci may generate the datasets indicative of a quantitative measure (e.g., indicative of a presence, absence, or relative amount) of sequences at each of the plurality of cancer-associated genomic loci in the cell-free biological sample. Assay readouts may comprise quantitative PCR (qPCR) values, digital PCR (dPCR) values, digital droplet PCR (ddPCR) values, fluorescence values, or normalized values thereof.

## EXAMPLES

### EXAMPLE 1: Selection of Methylated Regions for Detection of Multiple Cancer Types

**[0358]** To design a signature panel capable of detecting and distinguishing multiple types of cancers, regions of cfDNA that are methylated in various types of cancers and capable of being used to determine tissue of origin of a cancer type (tumor or cancerous cells) were identified. Two principles were used for designing a multi-cancer signature panel of methylated regions of DNA:

(i) identification of regions useful for screening for different cancer types including regions that may considered "pan-cancer" and methylated in more than one type of cancer; and

(ii) identification of regions useful for determining the tissue of origin of the tumor (TOO) including regions that are methylated or hypermethylated only in one cancer of interest and not in other cancer types or in subjects not having any cancer.

### TCGA and EPIC array data analysis

**[0359]** TCGA 450K array data were used for analysis. 450K methylation array raw idat files for 33 cancer types (including cancer and normal tissue data) were downloaded from the TCGA website. Beta values for each probe were calculated using the R package SeSAMe. Each region in the CpG dense light panel (CpGdv2) was assigned the average beta value of all probes overlapping the region. **Table 19** shows the number of cancer and normal tissue data obtained.

**Table 19**

| Symbol | Cancer type | # Cancer | # Normal | # Total |
|---|---|---|---|---|
| COAD | Colon adenocarcinoma | 314 | 39 | 353 |
| LIHC | Liver hepatocellular carcinoma | 380 | 50 | 430 |
| LUAD | Lung adenocarcinoma | 475 | 32 | 507 |
| LUSC | Lung squamous cell carcinoma | 370 | 42 | 412 |
| OV | Ovarian serous cystadenocarcinoma | 10 | 0 | 10 |
| PAAD | Pancreatic adenocarcinoma | 185 | 10 | 195 |
| PRAD | Prostate adenocarcinoma | 503 | 50 | 553 |
| READ | Rectum adenocarcinoma | 99 | 7 | 106 |

**[0360]** Public blood EPIC array data used for analysis was downloaded from GEO (Blood, GSE110555, 67 samples). The public blood data was generated on the EPIC array, so only probes that overlapped the TCGA 450K array data were used. Each region in the CpG dense

light panel was assigned a beta value similar to the procedure described above for the TCGA data.

**Univariate analysis**

**[0361]** Univariate AUCs for each region in the CpG dense light panel were calculated for cancer vs. normal tissue (for all cancers that had normal tissue data), and cancer vs. blood (for all cancers). Regions that had univariate AUC $\geq$0.9 for both the cancer vs. blood and cancer vs. normal tissue comparison were kept for downstream analyses. This resulted in a total of 3840 regions, adding up to 6349802 bp in size.

**Metilene analysis**

**[0362]** Metilene analysis was performed on 450K methylation array tissue data from the TCGA, excluding data from non-cancer samples. Probe beta values that had been normalized using the OpenSesame R pipeline were used. Differentially methylated regions (DMRs) were retained that had a q-value of 0.05 or less. The overlap of these regions with the CpG Dense panel were examined. Each CpG Dense region was annotated as detected by Metilene or not detected in each tissue type. This information was used to identify regions that were detected in a single tissue and could be used for tissue of origin detection vs. multiple tissues. This resulted in a total of 3498 regions, adding up to 4276029 bp in size.

***Overlap between the univariate analysis and metilene analysis***

**[0363]** ~2.2 Mb (1681 regions) overlapped between the univariate and metilene analysis. These regions were further used for downstream analysis and filtered based on overlap with the regions from the HMFC analysis of tissue TEM-seq data described later.

**[0364] FIG. 2** provides a heatmap of beta values of these 1681 regions that indicates that these regions may contain useful signal for determining tumor of origin as well. Different tumor types cluster into largely distinct groups. The heatmap shows clustering of beta values from the regions identified from the analysis. Colon adenocarcinoma (COAD) and Rectal adenocarcinoma (READ) clustered together. Lung squamous carcinoma (LUSC) and Lung adenocarcinoma (LUAD) formed largely two independent groups with a few samples that overlap. Total region size in this analysis was ~2.2Mb.

***Identifying Tissue of Origin Regions from TCGA Analysis***

**[0365]** For the 1681 regions from the TCGA analysis that overlapped the univariate and metilene analysis, a putative list of TOOs was defined having DMRs in only one cancer type. These regions were verified by performing univariate analysis for one vs. every other cancer type, and keeping regions that are concordant for tissue type between the metilene and univariate

analysis. Regions which had a univariate AUC ≥0.75 for the cancer were considered DMRs, whereas <0.65 AUC for every other cancer type were kept for the final putative TOO list from the TCGA analysis. This analysis resulted in 79 regions with a total size of 103,554 bp.

***Analysis of Tissue Methyl-seq Data***

**Data**

**[0366]** FF (flash frozen) tissue retrospective samples were obtained. DNA isolated therefrom was sequenced with methylation-sequence methods. **Table 20** shows the number of samples for each tissue sample obtained.

**Table 20**

| Tissue | # Samples |
|---|---|
| CRC | 63 |
| Liver | 14 |
| Lung | 24 (in duplicate) |
| Ovarian | 22 |
| Pancreatic | 29 |
| Prostate | 29 |
| Healthy plasma | 96 |

**Autosegmentation**

**[0367]** A modified version of the auto-segmentation pipeline was used to define reasonable region boundaries for each cancer type. Filtered and unfiltered bam files were created for each cancer type. Pickle files were created and input into a modified autosegmentation pipeline to identify regions that have methylation in cancer samples but little to no methylation in the healthy plasma samples.

**Hypermethylated Fragment Analysis in cancer vs. plasma models for feature selection**

**[0368]** Hypermethylated fragment analysis was used and summarized over the segmented regions for each cancer. To identify top features, hypermethylated fragment analysis was performed for cancer vs. plasma models using 5-fold CV with 5 reshuffles, keeping regions that were selected in at least 1-fold and have a mean effect size >$90^{th}$ percentile. This resulted in 845 regions with a total region size of 643185 bp.

**Hypermethylated Fragment Analysis in cancer vs. every other cancer model for putative TOO feature selection**

**[0369]** For each cancer type, regions that are hypermethylated in a cancer of interest but not in any other cancers were identified. To achieve this, hypermethylated fragment analysis was used, keeping regions selected in all 25-folds and a mean effect size the lesser of $100^{th}$ or the $99^{th}$ percentile value. This resulted in a total of 141 regions with a total size of 86,129 bp.

**Final multi-cancer panel design procedure**

**[0370]** Regions from the TCGA univariate analysis that overlapped both the metilene differentially methylated region analysis and methylated fragment tissue methyl-seq analysis were combined with the putative TOO regions identified either from the TCGA or methyl-seq tissue data analysis to obtain a multi-cancer signature panel. This resulted in a total of 417 methylated regions with a total size of 512,123 bp.

**[0371] FIG. 3** shows a heatmap of the regions included in the multi-cancer panel. The heatmap shows distinct separation between the different cancer types even with this smaller subset. The heatmap shows clustering of beta values from the regions identified from the analysis. Colon adenocarcinoma (COAD) and Rectal adenocarcinoma (READ) clustered together. Lung squamous carcinoma (LUSC) and Lung adenocarcinoma (LUAD) formed largely two independent groups with a few samples that overlapped.

## CLAIMS

WHAT IS CLAIMED IS:

1. A methylation signature panel characteristic of at least two cell proliferative disorders comprising:

one or more genomic regions selected from the group consisting of genomic regions in **Table 1**, wherein the one or more genomic regions are more methylated in a biological sample from a subject having a cell proliferative disorder or subtype thereof, and are less methylated in a biological sample from a subject not having the cell proliferative disorder or subtype thereof.

2. The methylation signature panel of claim 1, wherein the biological sample is a nucleic acid, DNA, RNA, or cell-free nucleic acid.

3. The methylation signature panel of claim 1, wherein the one or more genomic regions are non-coding regions, coding regions, non-transcribed regions, or regulator regions.

4. The methylation signature panel of claim 1, wherein the methylation signature panel comprises six or more genomic regions selected from the group consisting of genomic regions in **Table 1**.

5. The methylation signature panel of claim 1, wherein the one or more genomic regions selected from the group consisting of genomic regions in **Table 1** is associated with a type of cancer.

6. The methylation signature panel of claim 1, wherein the biological sample obtained from the subject having the cell proliferative disorder or subtype thereof is selected from the group consisting of body fluids, stool, colonic effluent, urine, blood plasma, blood serum, whole blood, isolated blood cells, cells isolated from the blood, and combinations thereof.

7. The methylation signature panel of claim 1, wherein the biological sample obtained from the subject not having the cell proliferative disorder or subtype thereof is selected from the group consisting of body fluids, stool, colonic effluent, urine, blood plasma, blood serum, whole blood, isolated blood cells, cells isolated from the blood, and combinations thereof.

8. The methylation signature panel of claim 1, wherein the cell proliferative disorder is selected from the group consisting of colorectal cell proliferation, prostate cell

proliferation, lung, breast cell proliferation, pancreatic cell proliferation, ovarian cell proliferation, uterine cell proliferation, liver cell proliferation, esophagus cell proliferation, stomach cell proliferation, and thyroid cell proliferation.

9. The methylation signature panel of claim 1, wherein the cell proliferative disorder is selected from the group consisting of colon adenocarcinoma, liver hepatocellular carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, ovarian serious cystadenocarcinoma, pancreatic adenocarcinoma, prostate adenocarcinoma, and rectum adenocarcinoma.

10. The methylation signature panel of claim 1, wherein the cell proliferative disorder is selected from the group consisting of stage 1 cancer, stage 2 cancer, stage 3 cancer, and stage 4 cancer.

11. The methylation signature panel of claim 1, wherein the signature panel comprises two or more genomic regions selected from the group consisting of genomic regions in **Table 1**, three or more genomic regions selected from the group consisting of genomic regions in **Table 1**, four or more genomic regions selected from the group consisting of genomic regions in **Table 1**, five or more genomic regions selected from the group consisting of genomic regions in **Table 1**, six or more genomic regions selected from the group consisting of genomic regions in **Table 1**, seven or more genomic regions selected from the group consisting of genomic regions in **Table 1**, eight or more methylated genomic regions selected from the group consisting of genomic regions in **Table 1**, nine or more genomic regions selected from the group consisting of genomic regions in **Table 1**, ten or more genomic regions selected from the group consisting of genomic regions in **Table 1**, eleven or more genomic regions in genomic regions in **Table 1**, twelve or more genomic regions selected from the group consisting of genomic regions in **Table 1**, or thirteen or more genomic regions selected from the group consisting of genomic regions in **Table 1**.

12. A methylation signature panel characteristic of a tissue of origin for at least two cell proliferative disorders comprising:

two or more genomic regions selected from the group consisting of genomic regions in **Tables 2-17**, wherein the two or more genomic regions are more methylated in a biological sample from a subject having a cell proliferative disorder or subtype thereof, and are less methylated in a biological sample from a subject not having the cell proliferative disorder or subtype thereof.

13. The methylation signature panel of claim 12, wherein the biological sample is a nucleic acid, DNA, RNA, or cell-free nucleic acid.

14. The methylation signature panel of claim 12, wherein the two or more genomic regions are non-coding regions, coding regions, non-transcribed regions, or regulator regions.

15. The methylation signature panel of claim 12, wherein the methylation signature panel comprises six or more genomic regions selected from the group consisting of genomic regions in **Tables 2-17**.

16. The methylation signature panel of claim 12, wherein the one or more genomic regions selected from the group consisting of genomic regions in **Tables 2-17** are associated with a type of cancer and a tumor tissue-of-origin.

17. The methylation signature panel of claim 12, wherein the biological sample obtained from the subject having the cell proliferative disorder or subtype thereof is selected from the group consisting of body fluids, stool, colonic effluent, urine, blood plasma, blood serum, whole blood, isolated blood cells, cells isolated from the blood, and combinations thereof.

18. The methylation signature panel of claim 12, wherein the biological sample obtained from the subject not having the cell proliferative disorder or subtype thereof is selected from the group consisting of body fluids, stool, colonic effluent, urine, blood plasma, blood serum, whole blood, isolated blood cells, cells isolated from the blood, and combinations thereof.

19. The methylation signature panel of claim 12, wherein cell proliferative disorder is selected from the group consisting of colorectal cell proliferation, prostate cell proliferation, lung cell proliferation, breast cell proliferation, pancreatic cell proliferation, ovarian cell proliferation, uterine cell proliferation, liver cell proliferation, esophagus cell proliferation, stomach cell proliferation, or thyroid cell proliferation.

20. The methylation signature panel of claim 12, wherein the cell proliferative disorder is selected from the group consisting of colon adenocarcinoma, liver hepatocellular carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, ovarian serious cystadenocarcinoma, pancreatic adenocarcinoma, prostate adenocarcinoma, and rectum adenocarcinoma.

21. The methylation signature panel of claim 12, wherein cell proliferative disorder is selected from the group consisting of stage 1 cancer, stage 2 cancer, stage 3 cancer, and stage 4 cancer.

22. The methylation signature panel of claim 12, wherein signature panel comprises three or more genomic regions selected from the group consisting of genomic regions in **Tables 2-17**, four or more genomic regions selected from the group consisting of genomic regions in **Tables 2-17**, five or more genomic regions selected from the group consisting of genomic regions in **Tables 2-17,** six or more genomic regions selected from the group consisting of genomic regions in **Tables 2-17**, seven or more genomic regions selected from the group consisting of genomic regions in **Tables 2-17**, eight or more genomic regions selected from the group consisting of genomic regions in **Tables 2-17**, nine or more genomic regions selected from the group consisting of genomic regions in **Tables 2-17**, ten or more genomic regions selected from the group consisting of genomic regions in **Tables 2-17**, eleven or more genomic regions in genomic regions in **Tables 2-17**, twelve or more genomic regions selected from the group consisting of genomic regions in **Tables 2-17**, or thirteen or more genomic regions selected from the group consisting of genomic regions in **Tables 2-17**.

23. The methylation signature panel of claim 12, wherein the at least two cell proliferative disorders comprise a combination selected from the group consisting of: colorectal cancer and prostate cancer; colorectal cancer and lung cancer; colorectal cancer and breast cancer; colorectal cancer and liver cancer; colorectal cancer and ovarian cancer; colorectal cancer and pancreatic cancer; prostate cancer and lung cancer; prostate cancer and breast cancer; prostate cancer and liver cancer; prostate cancer and ovarian cancer; prostate cancer and pancreatic cancer; lung cancer and breast cancer; lung cancer and liver cancer; lung cancer and ovarian cancer; lung cancer and pancreatic cancer; breast cancer and liver cancer; breast cancer and ovarian cancer; breast cancer and pancreatic cancer; liver cancer and ovarian cancer; liver cancer and pancreatic cancer; ovarian cancer and pancreatic cancer; colorectal cancer, prostate cancer, and lung cancer; colorectal cancer, prostate cancer, and breast cancer; colorectal cancer, prostate cancer, and liver cancer; colorectal cancer, prostate cancer, and ovarian cancer; colorectal cancer, prostate cancer, and pancreatic cancer; colorectal cancer, lung cancer, and breast cancer; colorectal cancer, lung cancer, and liver cancer; colorectal cancer, lung cancer, and ovarian cancer; colorectal cancer, lung cancer, and pancreatic cancer; colorectal

cancer, breast cancer, and liver cancer; colorectal cancer, breast cancer, and ovarian cancer; colorectal cancer, breast cancer, and pancreatic cancer; prostate cancer, liver cancer, and ovarian cancer; prostate cancer, liver cancer, and pancreatic cancer; prostate cancer, ovarian cancer, and pancreatic cancer; and colorectal cancer, prostate cancer, lung cancer, and breast cancer.

24. The methylation signature panel of claim 12, wherein the two or more genomic regions are selected from the group consisting of genomic regions in **Tables 2**, **3**, and **4**, and are associated with a colorectal cancer tissue of origin.

25. The methylation signature panel of claim 12, wherein the two or more genomic regions are selected from the group consisting of genomic regions in **Tables 5**, **6**, and **7**, and are associated with a liver cancer tissue of origin.

26. The methylation signature panel of claim 12, wherein the two or more genomic regions are selected from the group consisting of genomic regions in **Tables 8** and **9**, and are associated with a lung cancer tissue of origin.

27. The methylation signature panel of claim 12, wherein the two or more genomic regions are selected from the group consisting of genomic regions in **Tables 10**, **11**, and **12**, and are associated with an ovarian cancer tissue of origin.

28. The methylation signature panel of claim 12, wherein the panel of two or more genomic regions are selected from the group consisting of genomic regions in **Tables 13** and **14**, and are associated with a pancreatic cancer tissue of origin.

29. The methylation signature panel of claim 12, wherein the two or more genomic regions are selected from the group consisting of genomic regions in **Tables 15**, **16**, and **17**, and are associated with a prostate cancer tissue of origin.

30. A machine learning classifier capable of distinguishing a population of healthy subjects from subjects having a cell proliferative disorder, comprising:

a) sets of measured values representative of differentially-methylated genomic regions of the group consisting of differentially-methylated genomic regions in **Tables 1-17**, wherein the differentially-methylated genomic regions are associated with at least two cell proliferative disorders, where the measured values are obtained from methylation sequencing data from healthy subjects and subjects having a cell proliferative disorder,

b) wherein the measured values are used to generate a set of features corresponding to properties of the differentially-methylated genomic regions and where the features are analyzed using a machine learning or statistical model,

c) wherein the model provides a feature vector useful as a classifier capable of distinguishing a population of healthy subjects from subjects having a cell proliferative disorder.

31. The machine learning classifier of claim 30, wherein the sets of measured values describe characteristics of the methylated regions selected from the group consisting of: base wise methylation percent for CpG, CHG, CHH; the count or rate of observing fragments with different counts or rates of methylated CpGs in a region; conversion efficiency (100-mean methylation percent for CHH); hypomethylated blocks; methylation levels (global mean methylation for CPG, CHH, CHG, fragment length, fragment midpoint, and methylation levels in one or more genomic regions such as chrM, LINE1, or ALU); number of methylated CpGs per fragment; fraction of CpG methylation to total CpG per fragment; fraction of CpG methylation to total CpG per region; fraction of CpG methylation to total CpG in panel; dinucleotide coverage (normalized coverage of dinucleotide); evenness of coverage (unique CpG sites at lx and 10x mean genomic coverage (for S4 runs); mean CpG coverage (depth) globally; and mean coverage at CpG islands, CGI shelves, and CGI shores.

32. The machine learning classifier of claim 30, wherein machine learning classifier is capable of identifying a tissue of origin of a tumor in the subject.

33. The machine learning classifier of claim 30, wherein the classifier is loaded into a memory of a computer system, wherein the model is trained using training vectors obtained from training biological samples, a first subset of the training biological samples identified as having a cell proliferative disorder, and a second subset of the training biological samples identified as not having a cell proliferative disorder.

34. A machine learning classifier of claim 30, wherein the model is trained on a panel of predetermined methylated genomic regions associated with at least two cell proliferative disorders, and having pre-selected sensitivity and specificity for the different types of cell proliferative disorder to be detected using the panel.

35. The machine learning classifier of claim 30, wherein the at least two cell proliferative disorders are selected from the group consisting of colorectal cancer, breast cancer,

ovarian cancer, prostate cancer, lung cancer, pancreatic cancer, uterine cancer, liver cancer, esophagus cancer, stomach cancer, thyroid cancer, and bladder cancer.

36. The machine learning classifier of claim 30, wherein the machine learning classifier is tailored to provide a pre-selected sensitivity and a pre-selected specificity for each of the at least two cell proliferative disorders, wherein the at least two cell proliferative disorders are selected from the group consisting of colorectal cancer, breast cancer, ovarian cancer, prostate cancer, lung cancer, pancreatic cancer, uterine cancer, liver cancer, esophagus cancer, stomach cancer, thyroid cancer, and bladder cancer,

wherein the pre-selected sensitivity for a colorectal cancer associated classification panel is at least 70% sensitivity; the pre-selected specificity for a breast cancer associated classification panel is at least 70% specificity; the pre-selected specificity for an ovarian cancer associated classification panel is at least 90% specificity; the pre-selected specificity for a prostate cancer associated classification panel is at least 70% specificity; the pre-selected specificity for a lung cancer associated classification panel is at least 70% specificity; the pre-selected specificity for a pancreatic cancer associated classification panel is at least 90% specificity; the pre-selected specificity for a uterine cancer associated classification panel is at least 90% specificity; the pre-selected sensitivity for a liver cancer associated classification panel is at least 70% sensitivity; the pre-selected sensitivity for an esophagus cancer associated classification panel is at least 70% sensitivity; the pre-selected sensitivity for a stomach cancer associated classification panel is at least 70% sensitivity; the pre-selected specificity for a thyroid cancer associated classification panel is at least 70% specificity; and the pre-selected sensitivity for a bladder cancer associated classification panel is at least 70% sensitivity selected based on which cancer types are detected by the classification model.

37. A method for determining a methylation profile of a cell-free deoxyribonucleic acid (cfDNA) sample from a subject, comprising:

a) providing conditions for converting unmethylated cytosines to uracils in nucleic acid molecules of the cfDNA sample to produce a plurality of converted nucleic acid molecules;

b) contacting the plurality of converted nucleic acids with nucleic acid probes complementary to a pre-identified methylation signature panel characteristic of at least two cell proliferative disorders, wherein the methylation signature panel comprises

one or more genomic regions selected from the group consisting of genomic regions in **Tables 1-17** to enrich for sequences corresponding to the pre-identified methylation signature panel;

c) determining nucleic acid sequences of the plurality of converted nucleic acid molecules; and

d) aligning the nucleic acid sequences of the plurality of converted nucleic acid molecules to a reference nucleic acid sequence, thereby determining the methylation profile of the subject.

38. The method of claim 37, further comprising amplifying the plurality of converted nucleic acids.

39. The method of claim 38, wherein the amplifying comprises polymerase chain reaction (PCR).

40. The method of claim 37, further comprising preparing a nucleic acid sequencing library.

41. The method of claim 40, further comprising amplifying the plurality of converted nucleic acids, wherein the nucleic acid sequencing library is prepared prior to amplification.

42. The method of claim 37, further comprising determining the nucleic acid sequences of the converted nucleic acid molecules at a depth of greater than 1000x, greater than 2000x, greater than 3000x, greater than 4000x, or greater than 5000x.

43. The method of claim 37, wherein the reference nucleic acid sequence is at least a portion of a human reference genome.

44. The method of claim 37, wherein the methylation signature panel comprises three or more methylated genomic regions from the group consisting of methylated genomic regions in **Tables 1-17**, four or more methylated genomic regions from the group consisting of methylated genomic regions in **Tables 1-17**, five or more methylated genomic regions from the group consisting of methylated genomic regions in **Tables 1-17**, six or more methylated genomic regions from the group consisting of methylated genomic regions in **Tables 1-17**, seven or more methylated genomic regions from the group consisting of methylated genomic regions in **Tables 1-17**, eight or more methylated genomic regions from the group consisting of methylated genomic regions in **Tables 1-17**, nine or more methylated genomic regions from the group consisting of methylated genomic regions in **Tables 1-17**, ten or more methylated genomic regions from the group consisting of methylated genomic regions in **Tables 1-17**, eleven or more

methylated genomic regions from the group consisting of methylated genomic regions in **Tables 1-17**, twelve or more methylated genomic regions from the group consisting of methylated genomic regions in **Tables 1-17**, or thirteen or more methylated genomic regions from the group consisting of methylated genomic regions in **Tables 1-17**.

45. The method of claim 37, wherein the methylation profile is associated with a cell proliferative disorder and indicative of whether a subject has a cell proliferative disorder.

46. The method of claim 37, further comprising ligating a nucleic acid adapter comprising a unique molecular identifier to unconverted nucleic acids in the cfDNA sample prior to step a).

47. The method of claim 37, wherein the conditions for converting unmethylated cytosines to uracils in nucleic acid molecules of the cfDNA sample comprise chemical methods, enzymatic methods, or a combination thereof.

48. The method of claim 37, further comprising treating the cfDNA sample with a reagent selected from the group consisting of bisulfite, hydrogen sulfite, disulfite, and combinations thereof.

49. The method of claim 37, wherein the cfDNA sample obtained from the subject is selected from the group consisting of body fluids, stool, colonic effluent, urine, blood plasma, blood serum, whole blood, isolated blood cells, cells isolated from the blood, and combinations thereof.

50. The method of claim 37, further comprising applying a trained machine learning classifier to the methylation profile of the subject, wherein the trained machine learning classifier is trained to be capable of distinguishing between healthy subjects and subjects with a cell proliferative disorder to provide an output value associated with presence of a cell proliferative disorder, thereby detecting the presence or the absence of the cell proliferative disorder in the subject.

51. The method of claim 50, wherein the output value is at least 15%.

52. The method of claim 37, wherein the cell proliferative disorder is selected from the group consisting of stage 1 cancer, stage 2 cancer, stage 3 cancer, and stage 4 cancer.

53. A method for detecting a cell proliferative disorder in a subject comprising:

a) obtaining methylation sequencing information for a preselected panel of genomic regions associated with the presence of at least two different cell proliferative disorder tissue types from a nucleic acid sample from the subject,

b) applying the sequence information from the subject to a classification model trained on a preselected panel of genomic regions associated with the presence of the at least two cell proliferative disorder types, to identify the presence of a cell proliferative disorder, and if a cell proliferative disorder is detected, and

c) applying sequence information from the subject to a classification model trained on a preselected panel of genomic regions associated with associated with the presence of cell proliferative disorders in different tissue types to determine tissue of origin of the cell proliferative disorder in the subject.

54. A method for detecting a cell proliferative disorder in a subject, comprising

a) obtaining methylation sequencing information disorders from a nucleic acid sample from the subject for a preselected panel of genomic regions associated with at least two different cell proliferative disorders,

b) calculating a methylation profile of cfDNA in the sample corresponding to the preselected panel of predetermined methylated genomic regions associated with the at least two types of cell proliferative disorders, and

c) applying a machine learning classifier trained on a panel of predetermined methylated genomic regions associated with two or more types of cell proliferative disorder, and having pre-selected sensitivity and specificity for the different types of cell proliferative disorder to be detected using the panel.

55. The method of claim 53 or 54, wherein the different types of cell proliferative disorders are selected from the group consisting of colorectal cancer, breast cancer, ovarian cancer, prostate cancer, lung cancer, pancreatic cancer, uterine cancer, liver cancer, esophagus cancer, stomach cancer, thyroid cancer, or bladder cancer.

56. The method of claim 53 or 54, wherein the machine learning classifier is tailored to provide pre-selected sensitivity and specificity for the different types of cell proliferative disorder to be detected depending on needs of cancer diagnosis and confirmatory diagnosis for two or more cancers selected from the group consisting of colorectal cancer, breast cancer, ovarian cancer, prostate cancer, lung cancer, pancreatic cancer, uterine cancer, liver cancer, esophagus cancer, stomach cancer, thyroid cancer, and bladder cancer,

wherein the pre-selected sensitivity for a colorectal cancer associated classification panel is at least 70% sensitivity; the pre-selected specificity for a breast cancer associated classification panel is at least 70% specificity; the pre-selected specificity for an ovarian cancer associated classification panel is at least 90% specificity; the pre-selected specificity for a prostate cancer associated classification panel is at least 70% specificity; the pre-selected specificity for a lung cancer associated classification panel is at least 70% specificity; the pre-selected specificity for a pancreatic cancer associated classification panel is at least 90% specificity; the pre-selected specificity for a uterine cancer associated classification panel is at least 90% specificity; the pre-selected sensitivity for a liver cancer associated classification panel is at least 70% sensitivity; the pre-selected sensitivity for an esophagus cancer associated classification panel is at least 70% sensitivity; the pre-selected sensitivity for a stomach cancer associated classification panel is at least 70% sensitivity; the pre-selected specificity for a thyroid cancer associated classification panel is at least 70% specificity; and the pre-selected sensitivity for a bladder cancer associated classification panel is at least 70% sensitivity selected based on which cancer types are detected by the classification model.

57. A method for detecting a presence or an absence of a cell proliferative disorder in a subject, comprising:

a) providing conditions for converting unmethylated cytosines to uracils in nucleic acid molecules of a biological sample obtained or derived from the subject to produce a plurality of converted nucleic acid molecules;

b) contacting the plurality of converted nucleic acids with nucleic acid probes complementary to a pre-identified methylation signature panel of at least two differentially methylated regions selected from the group consisting of **Tables 1-17** to enrich for sequences corresponding to the signature panel;

c) determining nucleic acid sequences of the converted nucleic acid molecules;

d) aligning the nucleic acid sequences of the plurality of converted nucleic acid molecules to a reference nucleic acid sequence, thereby determining a methylation profile of the subject; and

e) applying a trained machine learning classifier to the methylation profile, wherein the trained machine learning classifier is trained to be capable of distinguishing between healthy subjects and subjects with a cell proliferative disorder to provide an

output value associated with presence of a cell proliferative disorder, thereby detecting the presence or the absence of the cell proliferative disorder in the subject.

58. A method for detecting a cell proliferative disorder in a subject, comprising:

a) providing conditions for converting unmethylated cytosines to uracils in nucleic acid molecules of a cfDNA sample to produce a plurality of converted nucleic acids;

b) amplifying converted nucleic acids with polymerase chain reaction;

c) probing the converted nucleic acids with nucleic acid probes complementary to a pre-identified methylation signature panel of at least two differentially methylated regions selected from the group consisting of **Tables 1-17** to enrich for sequences corresponding to the signature panel;

d) determining the nucleic acid sequence of the converted nucleic acid molecules at a depth of greater than 5000x, and

e) aligning the nucleic acid sequence of the converted nucleic acid molecules to a reference nucleic acid sequence for the pre-identified panel of CpG loci, to determine the methylation profile of the subject, and

f) analyzing the methylation profile using a machine learning model trained to be capable of distinguishing between healthy subjects and subjects with a cell proliferative disorder to provide an output value associated with presence of a cell proliferative disorder, thereby indicating the presence of a cell proliferative disorder in the subject.

59. The method of claim 57 or 58, wherein the biological sample obtained from the subject is selected from the group consisting of body fluids, stool, colonic effluent, urine, blood plasma, blood serum, whole blood, isolated blood cells, cells isolated from the blood, and combinations thereof.

60. The method of claim 57 or 58, wherein the method comprises applying the measured methylation signature panel from the subject against a database of measured methylation signature panels from normal subjects, wherein the database is stored on a computer system; determining that the subject has an increased risk of having a cell proliferative disorder by measuring a change of at least 15% in the methylation status of the methyl signature panel relative to methylation status from normal subjects.

61. The method of claim 57 or 58, wherein the cell proliferative disorder is selected from the group consisting of stage 1 cancer, stage 2 cancer, stage 3 cancer, and stage 4 cancer.

62. The method of claim 57 or 58, wherein the method detects pancreatic cancer and is performed in combination with detecting the presence or amount of CA19-9 protein in the biological sample.

63. The method of claim 57 or 58, wherein the method detects prostate cancer and is performed in combination with detecting the presence or amount of PSA protein in the biological sample.

64. A system comprising a machine learning model classifier for detecting a cell proliferative disorder, comprising:

a) a computer-readable medium comprising a classifier operable to classify subjects as having the cell proliferative disorder or not having the cell proliferative disorder based on a methylation signature panel of one or more genomic regions selected from the group consisting of genomic regions in **Tables 1-17**; and

b) one or more processors for executing instructions stored on the computer-readable medium.

65. The method of claim 64, wherein the system comprises the classifier loaded into a memory of a computer system, the machine learning model trained using training vectors obtained from training biological samples, a first subset of the training biological samples identified as having a cell proliferative disorder and a second subset of the training biological samples identified as not having a cell proliferative disorder.

66. The method of claim 64, wherein the classifier is provided in a system for detecting a cell proliferative disorder comprising:

a) a computer-readable medium comprising a classifier operable to classify the subjects based on a methylation signature panel described herein; and

b) one or more processors for executing instructions stored on the computer-readable medium.

67. The method of claim 64, wherein the system comprises a classification circuit that is configured as a machine learning classifier selected from the group consisting of a deep learning classifier, a neural network classifier, a linear discriminant analysis (LDA) classifier, a quadratic discriminant analysis (QDA) classifier, a support vector machine (SVM) classifier, a random forest (RF) classifier, a linear kernel support vector machine classifier, a first or second order polynomial kernel support vector machine classifier, a ridge regression classifier, an elastic net algorithm classifier, a sequential minimal

optimization algorithm classifier, a naive Bayes algorithm classifier, and principal component analysis classifier.

68. The method of claim 64, wherein the computer-readable medium is a non-transitory computer-readable medium comprising machine-executable code that, upon execution by one or more computer processors, implements any of the methods above or elsewhere herein.

69. The method of claim 64, wherein the system comprises one or more computer processors and computer memory coupled thereto, wherein the computer memory comprises machine-executable code that, upon execution by the one or more computer processors, implements any of the methods described herein.

70. A method for monitoring minimal residual disease in a subject previously treated for disease comprising:

determining a methylation profile as described herein as a baseline methylation state and repeating an analysis to determine the methylation profile at one or more predetermined time points wherein a change from baseline indicates a change in the minimal residual disease status at baseline in the subject.

71. The method of claim 70, wherein the minimal residual disease is selected from the group consisting of response to treatment, tumor load, residual tumor post-surgery, relapse, secondary screen, primary screen, and cancer progression.

72. The method of claim 70 for determining response to treatment.

73. The method of claim 70 for monitoring tumor load.

74. The method of claim 70 for detecting residual tumor post-surgery.

75. The method of claim 70 for detecting relapse.

76. The method of claim 70 for use as a secondary screen.

77. The method of claim 70 for use as a primary screen.

78. The method of claim 70 for monitoring cancer progression.

79. The method of claim 70, wherein the dataset is indicative of the presence or susceptibility of the cancer at a sensitivity of at least about 80%.

80. The method of claim 70, wherein the dataset is indicative of the presence or susceptibility of the cancer at a sensitivity of at least about 90%.

81. The method of claim 70, wherein the dataset is indicative of the presence or susceptibility of the cancer at a sensitivity of at least about 95%.

82. The method of claim 70, wherein the dataset is indicative of the presence or susceptibility of the cancer at a positive predictive value (PPV) of at least about 70%.

83. The method of claim 70, wherein the dataset is indicative of the presence or susceptibility of the cancer at a positive predictive value (PPV) of at least about 80%.

84. The method of claim 70, wherein the dataset is indicative of the presence or susceptibility of the cancer at a positive predictive value (PPV) of at least about 90%.

85. The method of claim 70, wherein the dataset is indicative of the presence or susceptibility of the cancer at a positive predictive value (PPV) of at least about 95%.

86. The method of claim 70, wherein the dataset is indicative of the presence or susceptibility of the cancer at a positive predictive value (PPV) of at least about 99%.

87. The method of claim 70, wherein the dataset is indicative of the presence or susceptibility of the cancer at a negative predictive value (NPV) of at least about 80%.

88. The method of claim 70, wherein the dataset is indicative of the presence or susceptibility of the cancer at a negative predictive value (NPV) of at least about 90%.

89. The method of claim 70, wherein the dataset is indicative of the presence or susceptibility of the cancer at a negative predictive value (NPV) of at least about 95%.

90. The method of claim 70, wherein the dataset is indicative of the presence or susceptibility of the cancer at a negative predictive value (NPV) of at least about 99%.

91. The method of claim 70, wherein the trained algorithm determines the presence or susceptibility of the cancer of the subject with an Area Under Curve (AUC) of at least about 0.90.

92. The method of claim 70, wherein the trained algorithm determines the presence or susceptibility of the cancer of the subject with an Area Under Curve (AUC) of at least about 0.95.

93. The method of claim 70, wherein the trained algorithm determines the presence or susceptibility of the cancer of the subject with an Area Under Curve (AUC) of at least about 0.99.

94. The method of claim 70, wherein the method further comprises presenting a report a graphical user interface of an electronic device of a user.

95. The method of claim 70, wherein the user is the subject, individual or patient.

96. The method of claim 70, wherein the method further comprises determining a likelihood of the determination of a presence or susceptibility of cancer in the subject, individual, or patient.

97. The method of claim 70, wherein the trained algorithm comprises a supervised machine learning algorithm.

98. The method of claim 70, wherein the supervised machine learning algorithm comprises a deep learning algorithm, a support vector machine (SVM), a neural network, or a Random Forest.

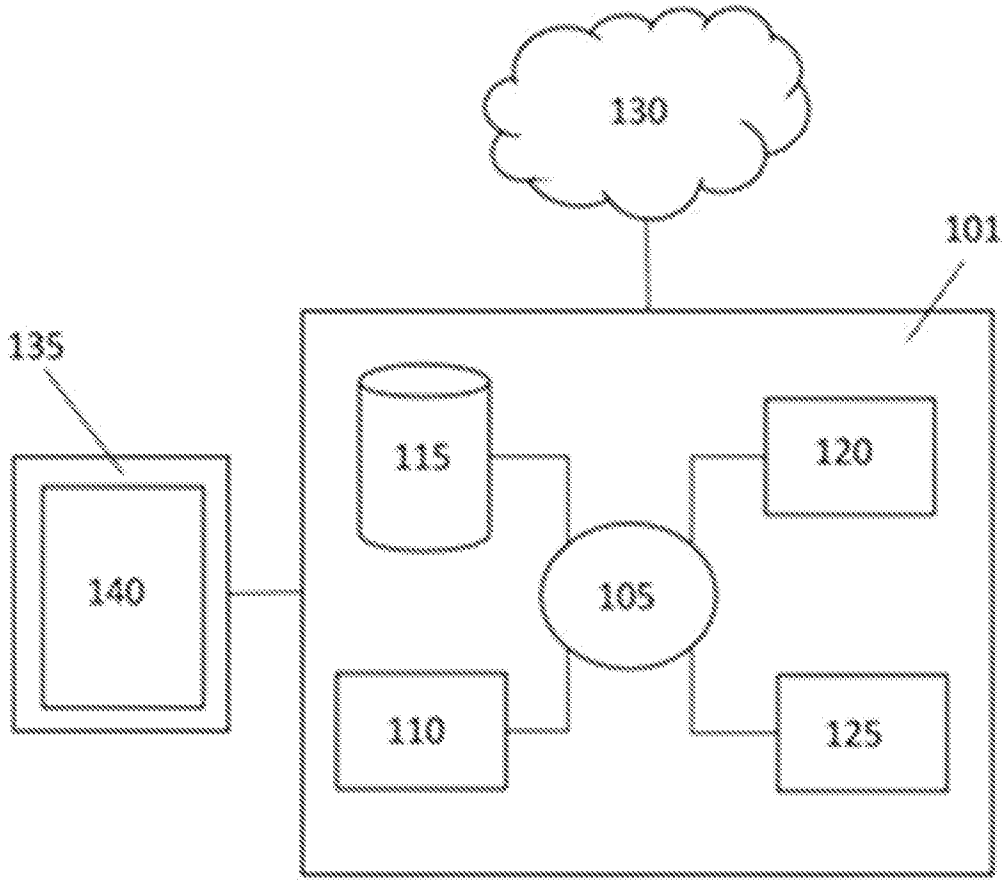
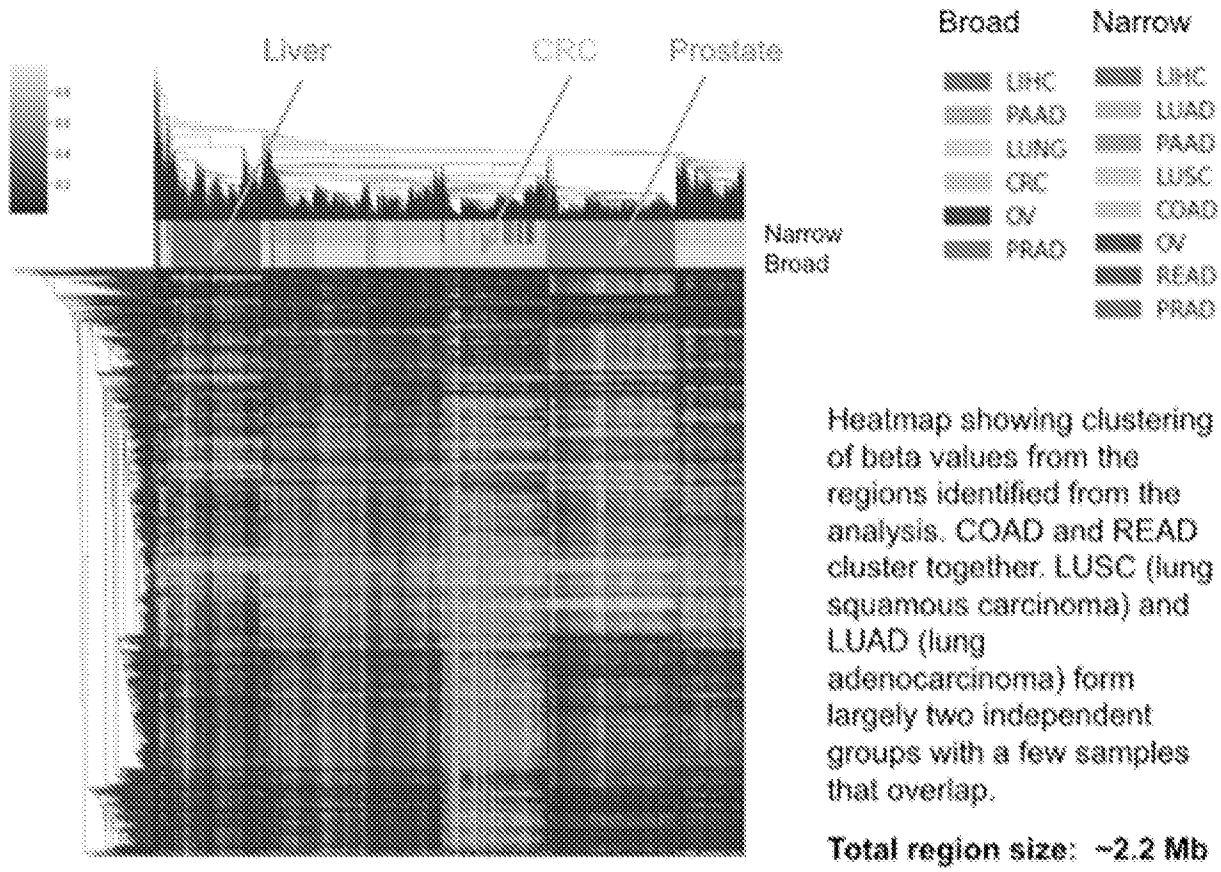



FIG. 1




Heatmap showing clustering of beta values from the regions identified from the analysis. COAD and READ cluster together. LUSC (lung squamous carcinoma) and LUAD (lung adenocarcinoma) form largely two independent groups with a few samples that overlap.

**Total region size: ~2.2 Mb**

**FIG. 2**

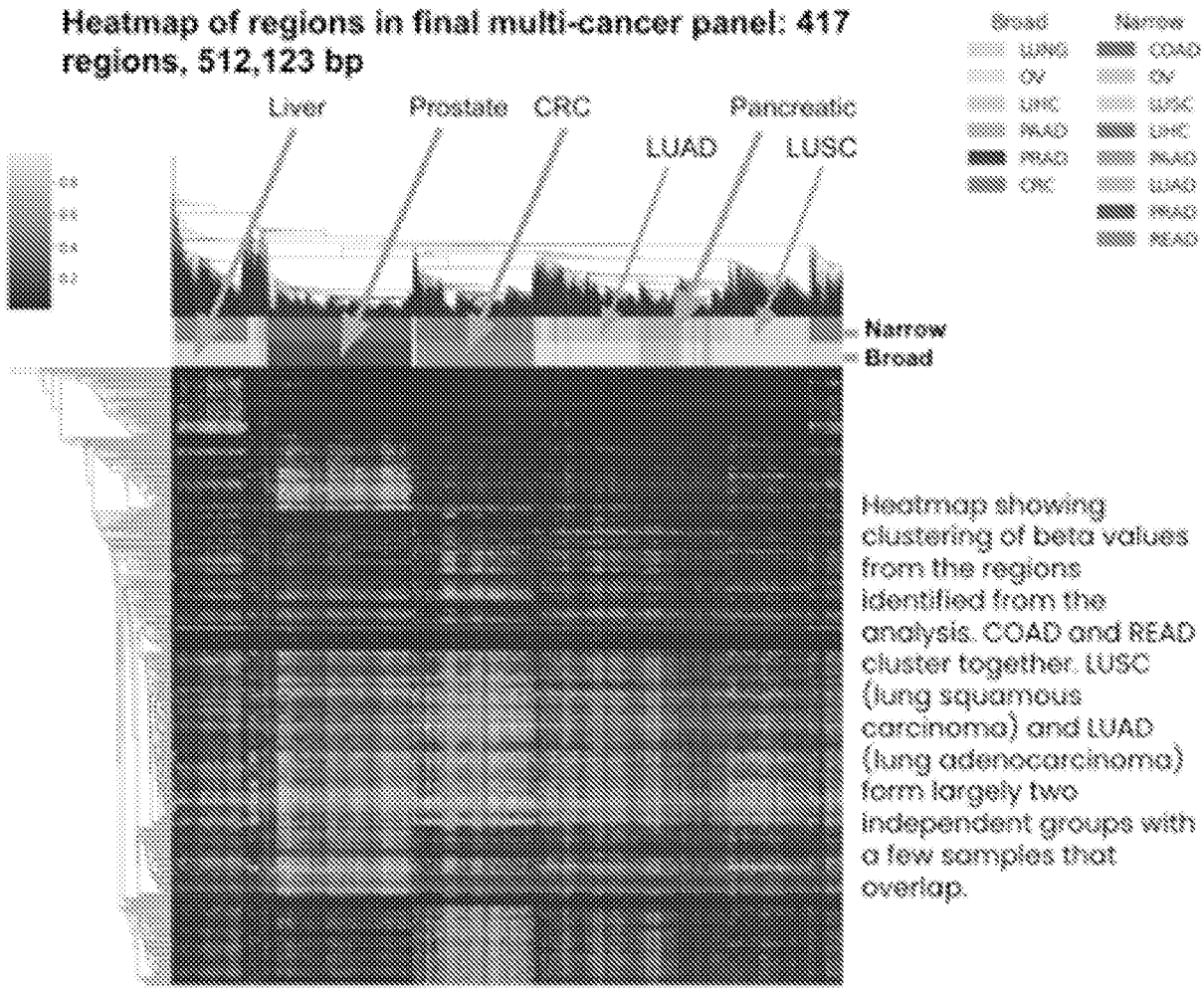

Heatmap of regions in final multi-cancer panel: 417 regions, 512,123 bp
Liver
Prostate
CRC
LUAD
Pancreatic
LUSC
Narrow
Broad
Heatmap showing clustering of beta values from the regions identified from the analysis. COAD and READ cluster together. LUSC (lung squamous carcinoma) and LUAD (lung adenocarcinoma) form largely two independent groups with a few samples that overlap.

FIG. 3

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 22/21662

A. CLASSIFICATION OF SUBJECT MATTER

IPC - C12Q 1/68, C12Q 1/6876, G16B 5/20, G16B 40/00 (2022.01)

CPC - C12Q 1/6869, C12Q 1/6886, G16B 5/20, G16B 20/00, G16B 40/20, C12Q 2600/154

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)
See Search History document

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| X | WO 2019/195268 A2 (GRAIL, INC.) 10 October 2019 (10.10.2019); abstract; para [0006]-[0007], [0027], [00169], [00176], [00204], [00246], [00282], [00287], [00290]-[00292], [00298], [00324], [00329], [00331],[00335], [00338], [00341], [00344]; Table 1 | 1-10, 30-36, 64-69 |
| A | US 2018/0119137 A1 (DRIVER, INC.) 03 May 2018 (03.05.2018); entire document | 1-10, 30-36, 64-69 |

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

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 08 July 2022 | AUG 02 2022 |
| Name and mailing address of the ISA/US<br>Mail Stop PCT, Attn: ISA/US, Commissioner for Patents<br>P.O. Box 1450, Alexandria, Virginia 22313-1450<br>Facsimile No. 571-273-8300 | Authorized officer<br>Kari Rodriquez<br>Telephone No. PCT Helpdesk: 571-272-4300 |

Form PCT/ISA/210 (second sheet) (July 2019)

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 22/21662

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

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

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

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

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

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

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

---See Supplemental Box ---

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

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

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

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-10, 30-36, 64-69 limited to chr4 1410251-1411075

**Remark on Protest**

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

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

☐ No protest accompanied the payment of additional search fees.

Box V.2 Citations and Explanations

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be searched, the appropriate additional search fees must be paid.

Groups I+: Claims 1-36, 64-69, drawn to a methylation signature panel characteristic of at least two cell proliferative disorders comprising one or more genomic regions that re more methylated in a biological sample from a subject having a cell proliferative disorder. The composition will be searched to the extent that the first named genomic region encompasses chr4 1410251-1411075 (see instant Specification, Table 1). This first named invention has been selected based on the guidance set forth in section 10.54 of the PCT International Search and Preliminary Examination Guidelines. It is believed that claims 1-10, 30-36, 64-69 encompass this first named invention, and thus these claims will be searched without fee to the extent that they encompass chr4 1410251-1411075. Additional genomic region(s) will be searched upon the payment of additional fees. Applicants must specify the claims that encompass any additionally elected genomic region(s). Applicants must further indicate, if applicable, the claims which encompass the first named invention, if different than what was indicated above for this group. Failure to clearly identify how any paid additional invention fees are to be applied to the "+" group(s) will result in only the first claimed invention to be searched. An exemplary election would be two genomic regions encompassing chr13 93226527-93229026 and chr20 8131515-8134057 (see instant Specification, Table 1) (Claims 12-21, 23-24, 30-36, 64-69).

Groups II+: Claims 37-63, drawn to a method for determining a methylation profile of a cell-free deoxyribonucleic acid (cfDNA) sample from a subject, wherein the methylation signature panel characteristic of at least two cell proliferative disorders, and comprises one or more genomic regions. Group II+ will be searched upon payment of additional fees. The method may be searched, for example, to the extent that the first named genomic region encompasses chr4 1410251-1411075 (see instant Specification, Table 1), for an additional fee and election as such. It is believed that claims 37-43, 45-56 read on this exemplary invention. Additional genomic region(s) will be searched upon the payment of additional fees. Applicants must specify the claims that encompass any additionally elected genomic region(s). Failure to clearly identify how any paid additional invention fees are to be applied to the"+" group(s) will result in only the first claimed invention to be searched. Another exemplary election would be two genomic regions encompassing chr13 93226527-93229026 and chr20 8131515-8134057 (see instant Specification, Table 1) (Claims 37-63).

Group III: Claims 70-98, drawn to a method for monitoring minimal residual disease in a subject previously treated for disease.

The inventions listed as Groups I+, II+ and III do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

Special Technical Features

Groups I+ include the special technical feature of a composition which differs from the special technical feature of a method, as disclosed by Groups II+ and III.

Groups II+ include the special technical feature of a method for determining a methylation profile of a sample from a subject, wherein the methylation signature panel characteristic of at least two cell proliferative disorders, not required by Group III.

The inventions of Groups I+ and II+ each include the special technical feature of a specific genomic region, not required by any of the other inventions of Groups I+ and II+.

Common Technical Features

The inventions of Groups I+, II+ and III share the technical feature of determining a methylation profile.

The inventions of Groups I+ and II+ share the technical feature of a methylation signature panel characteristic of at least two cell proliferative disorders comprising one or more genomic regions.

However, these shared technical features do not represent a contribution over prior art in view of WO 2019/195268 A2 to Grail, Inc. (hereinafter "Grail").

Grail discloses (instant claim 1) a methylation signature panel characteristic of at least two cell proliferative disorders (abstract - "The present description provides a cancer assay panel for targeted detection of cancer-specific methylation patterns"; para [00329] - "Cancer assay panels comprising probes targeting the selected genomic regions were generated. Specifically, the panels were designed to detect the presence and/or stage of cancer generally (i.e., vs non-cancer) or a specific cancer type as listed below: Table 1 : Pan-cancer #1; Table 2: Blood cancer #1....") comprising: one or more genomic regions (para [0007] - "the present description provides cancer assay panels (e.g., bait sets) for detecting cancer and various tissue or origins by detecting methylation patterns of targeted genomic regions") selected from the group consisting of genomic regions in Table 1 (para [00331] - "Generally, a probe can be designed to overlap any of the CpG sites included within the start/stop ranges of any of the targeted regions (e.g., anomalous fragments).....Table 1), wherein the one or more genomic regions are more methylated in a biological sample from a subject having a cell proliferative disorder or subtype thereof, and are less methylated in a biological sample from a subject not having the cell proliferative disorder or subtype thereof (Table 1, pg 144, (chr15, 26107640-2610786, Hyper); (chr15, 27018363-27018436, Hyper); (chr15, 27216396-27216420, Hyper, GABRG3)).

---See Supplemental Box ---

Box No. III Observations where unity of invention is lacking

Grail discloses (instant claim 37) a method for determining a methylation profile of a cell-free deoxyribonucleic acid, cfDNA, sample from a subject (para [0006] - "Targeted detection of methylation patterns specific to cancer or tissue of origin, i.e., the organ, organ group, body region or cell type that the cancer arises or origins from, using cell-free DNA (cfDNA) fragments can make early detection of cancer possible by providing a cost-effective and non-invasive method for analyzing information relevant to cancer classification"), comprising:
a. providing conditions for converting unmethylated cytosines to uracils in nucleic acid molecules of the cfDNA sample to produce a plurality of converted nucleic acid molecules (para [0027] - "Further disclosed herein are methods for providing sequence information informative of a presence or absence of cancer, comprising the steps of obtaining a test sample comprising a plurality of cfDNA test molecules; processing the cfDNA test molecules, thereby obtaining bisulfite-converted test fragments");
b. contacting the plurality of converted nucleic acids with nucleic acid probes complementary to a pre-identified methylation signature panel characteristic of at least two cell proliferative disorders (para [0029] - "the assay panel comprises a plurality of polynucleotide probes, wherein each of the polynucleotide probes is configured to hybridize to a bisulfite-converted fragment obtained from processing of cfDNA molecules wherein each of the cfDNA molecules corresponds to or is derived from one or more genomic regions selected from one of Tables 1, 12, 13, 14, and 15, wherein the cancer classification is a presence or absence of cancer or a stage of cancer"; para [00329] - "Cancer assay panels comprising probes targeting the selected genomic regions were generated. Specifically, the panels were designed to detect the presence and/or stage of cancer generally (i.e., vs non-cancer) or a specific cancer type as listed below: Table 1 : Pan-cancer #1; Table 2: Blood cancer #1....."), wherein the methylation signature panel comprises one or more genomic regions selected from the group consisting of genomic regions in Table 1 to enrich for sequences corresponding to the pre-identified methylation signature panel (para [00331] - "Generally, a probe can be designed to overlap any of the CpG sites included within the start/stop ranges of any of the targeted regions (e.g., anomalous fragments).....Table 1");
c. determining nucleic acid sequences of the plurality of converted nucleic acid molecules (para [00176] - "a cancer assay panel is designed to enrich nucleic acids derived from differentially methylated genomic regions in cancerous samples identified based on bisulfite sequencing data generated from the cfDNA from cancer and non-cancer individuals"); and
d. aligning the nucleic acid sequences of the plurality of converted nucleic acid molecules to a reference nucleic acid sequence, thereby determining the methylation profile of the subject (para [0280] - "the sequence reads may be aligned to a reference genome using known methods in the art to determine alignment position information").

As said technical features were known in the art at the time of the invention, these cannot be considered special technical features that would otherwise unify the groups.

Groups I+, II+ and III therefore lack unity under PCT Rule 13 because they do not share a same or corresponding special technical feature.

Note, Claims 65-69 which depend from "the method of claim 64", are objected to, because Claim 64 does not disclose "a method". Claims 65-69 are reconstrued as though depending from "the system of claim 64".