

CELL-FREE DNA HYDROXYMETHYLATION PROFILES IN THE EVALUATION OF PANCREATIC LESIONS

TECHNICAL FIELD

[0001] The present invention relates generally to epigenetic analysis, and more particularly relates to combined workflow methods for obtaining multiple types of information from a single biological sample. The invention finds utility in the fields of genomics, medicine, diagnostics, and epigenetic research.

BACKGROUND

[0002] Translational research using genomic and proteomic technologies has provided novel molecular insights into the pathogenesis of pancreatic cancer and the biology of the local tumor microenvironment, but has yet to yield robust diagnostic biomarkers to impact early diagnosis of a disease. This is reflected by a very low overall 5-year survival rate of 8.5%; see "Cancer Stat Facts: Pancreas Cancer" (National Cancer Institute Surveillance, Epidemiology, and End Results Program, 2017), retrieved on October 16, 2017 from seer.cancer.gov/statfacts/html/pancreas.html. Pancreatic cancer often presents late and has few symptoms, at which point only 10% to 20% of patients are eligible for surgical resection.

[0003] The pancreas consists of acinar cells, ductal cells, centro-acinar cells, endocrine islets, and stellate cells. The majority of pancreatic cancers are adenocarcinomas, with pancreatic ductal adenocarcinoma (PDAC) and its variants accounting for more than 90% of all pancreatic malignancies (Tempero et al. (2017) *Journal of the Comprehensive Cancer Network* 15(8): 1028- 1060), with the next most common pathology being neuroendocrine tumors, followed by colloid carcinomas, solid-pseudopapillary tumors, acinar cell carcinomas, and pancreatoblastomas (Kleef et al. (2016), *Nature Reviews Disease Primers: Pancreatic Cancer* 2: 1-22). Tobacco smoking confers a two- to three-fold higher risk of pancreatic cancer and also demonstrates a dose-risk relationship, while contributing to approximately 15 to 30% of cases (*ibid.*), with smokers diagnosed 8 to 15 years younger than non-smokers (Anderson et al. (2012) *Am. J. Gastroenterol* 107(11):1730-39; Maisonneuve et al. (2010) *Dig Dis* 28(405):645-56). A family history of pancreatitis is contributory in approximately 10% of cases, and germline mutations in genes such as BRCA2, BRCA1,

CDKN2A, ATM, STK11, PRSS1, MLH1 and PALB2 are also associated with pancreatic cancer with variable penetrance (Kleef, *supra*).

[0004] Age is a significant risk factor for pancreatic cystic lesions (PCLs) and pancreatic cancer. Zerboni et al. (2016) *Abstracts/Pancreatology* 16:S104 (Abstract ID: 1665) did a meta-analysis of 10 studies showing an overall prevalence of PCLs of 11%, with a higher rate of 16% in studies investigating subjects with a mean age greater than 55 years old. Studies using modern imaging technologies such as Magnetic Resonance Imaging (MRI) with contrast medium and cholangiopancreatography (MRCP) reported a significantly higher pooled prevalence of PCLs at 26% of subjects. Other known risk factors include, without limitation, diabetes mellitus, chronic pancreatitis, and obesity.

[0005] The management of PDAC presents physicians with challenges along the entire clinical spectrum, including early detection in high risk individuals, early diagnosis of patients with symptoms or imaging findings, prognostication of outcomes, and prediction of therapeutic responsiveness, which collectively have engendered intense research efforts to identify and validate biomarkers with sufficient clinical performance metrics to improve decision algorithms. Current guidelines in PDAC management are primarily limited to two biomarker recommendations: carbohydrate antigen 19-9 (CA19-9 or sialyl Lewis antigen) and carcinoembryonic antigen (CEA). CA19-9 is relied upon to guide surgery decisions, use of adjuvant therapy, or the detection of post-operative tumor recurrence, with the recognition that 10% of the population does not secrete the antigen. See Swords et al. (2016) *Onco Targets Ther* 9: 7459-67 and U.S. Patent No. 8,632,98. Furthermore, the restricted sensitivity and specificity of CA 19-09 as a biomarker for pancreatic cancer suggests limited diagnostic potential. CEA levels are assessed in pancreatic cyst fluid and then combined with imaging and clinical parameters to distinguish mucinous and non-mucinous cysts in order to mitigate risk (Fonseca et al. (2018) *Pancreas* 47(3): 272-79; Elta et al. (2018) *Am. J. Gastroenterology* 113: 464-79). However, CEA level does not correlate with the extent of disease (Schlieman et al. (2003) *Arch Surg.* 138(9): 951-56). Furthermore, while both tumor markers, if elevated, are useful in following patients with known disease, neither CA19-9 nor CEA has the sensitivity and specificity needed for use in screening patients to detect pancreatic cancer.

[0006] Molecular analyses of pancreatic cancer genomes reveal activating mutations in KRAS and inactivation of CDKN2A, TP53 and SMAD4, either through point mutation or copy number changes at >50% population frequency (Blankin et al. (2012) *Nature* 491(7424): 399-405; Waddell et al. (2015) *Nature* 518(7540):495-501; Jones et al. (2008)

Science 321(5897):1801-06); however, much mutational heterogeneity exists, rendering this subset of genes inefficient in diagnosing patients. Molecular subtyping of pancreatic tumors using mutational-based data (Waddell (2015), *supra*) or gene expression signatures (REF), have not yet seen clinical application.

[0007] There remains an unmet and pressing need in the art for improved methods of detecting, diagnosing, predicting, assessing, treating, and monitoring pancreatic cancer, particularly PDAC. An ideal method would be reliable and non-invasive, optimally enabling analysis of tumor, microenvironment, pancreas, and immune cell DNA to identify genetic and epigenetic information that correlates with PDAC or an aspect thereof.

SUMMARY OF THE INVENTION

[0008] Tumor and normal cell DNA is released into the bloodstream, and a cell-free DNA (cfDNA) sample extracted therefrom can be analyzed with respect to genetic and epigenetic signatures. Epigenetic signatures include, by way of example, DNA methylation, i.e., the conversion of cytosine to 5-methylcytosine (5mC), and DNA hydroxymethylation, the oxidation of 5mC to 5-hydroxymethylcytosine (5hmC), mediated in the mammalian genome by the TET (Ten-Eleven Translocation) family of enzymes. Such signatures may come from cells that are normal, or from a tumor, the tumor microenvironment, the affected organ, or the immune system, all of which may change in response to health conditions such as in the case of pancreatic cancer.

[0009] The present invention is predicated on the discovery of a set of hydroxymethylation biomarkers that in combination with one or more clinical parameters and optionally one or more other types of biomarkers and/or patient-specific risk factors, exhibits a hydroxymethylation level that correlates in some way with pancreatic cancer, particularly PDAC or another exocrine pancreatic cancer. In some embodiments, the invention enables the determination of:

[00010] (a) the risk that a pancreatic lesion observed with an imaging scan, i.e., an identified pancreatic lesion, is cancerous;

[00011] (b) the risk that an identified noncancerous pancreatic lesion will become cancerous;

[00012] (c) the likelihood that a particular therapy for treating a subject with pancreatic cancer will be effective;

- [00013] (d) the risk that a subject without an identified pancreatic lesion will, at some point, develop a pancreatic lesion, as well as
- [00014] (e) the risk of that lesion becoming cancerous.
- [00015] Observing changes in the biomarker set over time can provide (or in some cases confirm) additional information such as:
- [00016] (f) the effectiveness of a therapy a subject is undergoing in connection with an identified pancreatic lesion;
- [00017] (g) an increase or decrease in the risk that an identified pancreatic lesion will develop into cancer;
- [00018] (h) an increase or decrease in the likelihood that a subject without an observed pancreatic lesion will develop a pancreatic lesion, and
- [00019] (i) the risk of that lesion becoming cancerous; and
- [00020] (j) a change in an identified pancreatic lesion, including (j-1) a change in the size of a pancreatic lesion, (j-2) a change in the stage of a cancerous pancreatic lesion, (j-3) a change in the grade of a cancerous pancreatic lesion; (j-4) a change in the degree of invasiveness of a cancerous pancreatic lesion; and (j-5) the change from a local or regionalized invasive cancerous pancreatic lesion to a metastatic pancreatic cancer; as well as (j-6) the identification or confirmation of the pancreas as the primary tissue of origin in a cancer first identified through metastasis (i.e., initially a cancer of unknown origin).
- [00021] It will thus be appreciated that the methods herein remain useful after surgical resection of a pancreatic lesion, in the context of monitoring post-surgical changes such as the development of additional lesions or the effectiveness of a post-surgical therapy (e.g., radiation, chemotherapy, other pharmacotherapy, etc.) However, of primary importance herein is the evaluation of an identified pancreatic lesion as more or less likely to be cancerous or to become cancerous. Equally important is the ability to identify a likely cancerous lesion at an early stage. These features of the invention in turn enable significant advances in the field, including the treatment of pancreatic cancer before the cancer has advanced or metastasized as well as a reduction in unnecessary surgery, i.e., removal of benign lesions.
- [00022] In one embodiment of the invention, a method is provided for evaluating the risk that an identified pancreatic lesion in a patient is cancerous, the method comprising: (a) obtaining a cell-free DNA sample from the patient; (b) enriching for hydroxymethylated DNA in the sample; (c) quantifying the nucleic acids in the enriched sample that map to each

of a plurality of selected loci in a reference hydroxymethylation profile, wherein each selected locus comprises a hydroxymethylation biomarker; (d) comparing, at each locus, the hydroxymethylation level of the sample with the hydroxymethylation level in the reference profile, to ascertain differences in hydroxymethylation levels between the sample and the reference profile for each biomarker; and (e) calculating an index value representing the risk that the pancreatic lesion is cancerous from the comparison in step (d) combined with at least one additional parameter correlated with the risk that an individual has pancreatic cancer. The additional parameter may be a clinical parameter, an additional type of biological marker (i.e., a biological marker not related to hydroxymethylation), or a combination thereof.

[00023] The selected loci that serve as hydroxymethylation biomarkers herein comprise loci selected for their relevance to pancreatic cancer, particularly an exocrine pancreatic cancer such as PDAC. By "relevance" is meant that a hydroxymethylation biomarker locus, alone or in combination with one or more other hydroxymethylation biomarker loci, tends to exhibit an increase or decrease in hydroxymethylation in a manner that correlates with the risk, presence, absence, type, size, stage, invasiveness, grade, location, diagnosis, prognosis, outcome, and/or or likelihood of treatment responsiveness of pancreatic cancer, including determinations (a) through (j) above. The reference hydroxymethylation profile is a data set representing the hydroxymethylation level of each of a plurality of hydroxymethylation biomarkers, where the data set is a composite of hydroxymethylation profiles of a plurality of individuals having at least one shared characteristic.

[00024] It should be noted that some of the individual hydroxymethylation biomarkers disclosed herein may not have significant individual significance in the evaluation of a pancreatic lesion, but when used in combination with other hydroxymethylation biomarkers disclosed herein and clinical parameters impacting on the evaluation and monitoring of a pancreatic lesion, optionally further combined with one or more other types of biomarkers and/or patient-specific risk factors, become significant in discriminating as a method of the invention requires, e.g., between a subject who has pancreatic cancer and a subject who does not have pancreatic cancer, or between a subject who is likely to develop pancreatic cancer and a subject who is not likely to develop pancreatic cancer, etc. The methods of the present invention provide an improvement over currently available methods of evaluating the risk that a subject has pancreatic cancer or is likely to develop pancreatic cancer, by using the biomarkers defined herein.

[00025] In one aspect of the embodiment, a focused reference profile can be used to improve the accuracy of the above method. That is, different types of reference hydroxymethylation profiles may be constructed from different population groups, and an appropriate reference profile can then be selected for the evaluation of a particular patient. For patients who have chronic inflammation of the pancreas, i.e., chronic pancreatitis, a narrowed, or focused, reference profile generated from a set of individuals with chronic pancreatitis would be selected. Another focused reference profile might be constructed from a set of individuals who are diabetic, or obese, or cigarette smokers, and used in the evaluation of patients who are diabetic, obese, or smokers, respectively. These focused reference profiles can also be used in combination, depending on the attributes of the patient undergoing evaluation.

[00026] In another aspect, the cell-free DNA sample is extracted from a blood sample obtained from the patient. In another aspect, the cell-free DNA sample is extracted from a sample of pancreatic cyst fluid obtained from the patient.

[00027] In an additional aspect of the embodiment, step (b) comprises ligating adapters onto the DNA, functionalizing 5hmC residues in the DNA with an affinity tag that allows selective capture of tagged cfDNA, and removing the tagged cfDNA from the sample. The affinity tag may be a biotin moiety, in which case the functionalization of the 5hmC residues comprises biotinylation. The biotinylated cfDNA may then be captured using a solid support having a surface functionalized with a biotin-binding protein such as avidin or streptavidin. Step (b) may then further comprise amplifying the cfDNA without releasing the captured cfDNA from the support, to give a plurality of amplicons; sequencing the amplicons; and quantifying the nucleic acids that map to the reference loci from the sequence reads.

[00028] In another embodiment, a method is provided for monitoring a patient who has an identified pancreatic lesion, i.e., a lesion identified in an imaging scan. As with the preceding embodiment, the method is a non-invasive way of enabling the practitioner to identify changes in a previously identified pancreatic lesion and thereby determine, for example, whether the lesion is progressing toward cancer. The method comprises:

[00029] (a) obtaining an initial cell-free DNA sample from the patient;

[00030] (b) enriching for hydroxymethylated DNA in the initial sample;

[00031] (c) quantifying the nucleic acids in the enriched initial sample that map to each of a plurality of selected loci in a reference hydroxymethylation profile, wherein each selected locus comprises a hydroxymethylation biomarker;

[00032] (d) comparing, at each locus, the hydroxymethylation level of the enriched cell-free DNA in the initial sample with the hydroxymethylation level in the reference profile, to ascertain differences in hydroxymethylation levels between the sample and the reference profile for each biomarker;

[00033] (e) generating an initial hydroxymethylation profile for the patient comprising the hydroxymethylation level of the enriched cell-free DNA in the initial sample, at each locus;

[00034] (f) repeating steps (a) through (c) at a later time with a subsequent cell-free DNA sample obtained from the patient;

[00035] (g) generating a subsequent hydroxymethylation profile for the patient comprising the hydroxymethylation level of the enriched cell-free DNA in the subsequent sample, at each locus; and

[00036] (h) comparing, at each locus, the hydroxymethylation level of the enriched cell-free DNA in the subsequent sample to the hydroxymethylation level of the enriched cell-free DNA in the initial sample, to ascertain a change in the pancreatic lesion.

[00037] In the context of ongoing assessment, steps (f) through (h) are repeated at selected time intervals throughout an extended monitoring period.

[00038] The change in the pancreatic lesion is thus determined by a change in the patient's hydroxymethylation profile over time, at a plurality of hydroxymethylation biomarker loci, optimally in combination with one or more other risk factors or clinical parameters. The change in the lesion may be, for example, a change in size, a change in grade, a change in shape, a change in lymph node involvement, a change in invasiveness, or two or more of any of the foregoing.

[00039] In a related embodiment, the invention provides a method for managing a patient with a pancreatic lesion identified in an imaging scan, the method comprising:

[00040] (a) obtaining an initial cell-free DNA sample from the patient;

[00041] (b) enriching for hydroxymethylated DNA in the sample;

[00042] (c) quantifying the nucleic acids in the enriched initial sample that map to each of a plurality of selected loci in a reference hydroxymethylation profile, wherein each selected locus comprises a hydroxymethylation biomarker;

[00043] (d) comparing, at each locus, the hydroxymethylation level of the enriched cell-free DNA in the initial sample with the hydroxymethylation level in the reference profile, to ascertain differences in hydroxymethylation levels between the sample and the reference profile for each biomarker;

- [00044] (e) generating an initial hydroxymethylation profile for the patient comprising the hydroxymethylation level of the enriched cell-free DNA in the initial sample, at each locus;
- [00045] (f) repeating steps (a) through (c) at a later time with a subsequent cell-free DNA sample obtained from the patient;
- [00046] (g) generating a subsequent hydroxymethylation profile for the patient comprising the hydroxymethylation level of the enriched cell-free DNA in the subsequent sample, at each locus;
- [00047] (h) comparing, at each locus, the hydroxymethylation level of the enriched cell-free DNA in the subsequent sample to the hydroxymethylation level of the enriched cell-free DNA in the initial sample, to ascertain a change in the pancreatic lesion; and
- [00048] (f) based on the comparison in step (e), determining whether to treat the patient.
- [00049] Steps (a) through (h) of the method may be repeated at selected time intervals within the context of an ongoing monitoring period.
- [00050] If the changes in the patient's hydroxymethylation profile at a plurality of hydroxymethylation biomarker loci provides evidence of a change in the pancreatic lesion that, in the practitioner's opinion, warrants treatment, the treatment itself may be selected based on the change in the in the patient's hydroxymethylation profile at one or more of the selected loci. Treatment may involve radiation therapy, chemotherapy, other pharmacotherapy, surgical resection of the lesion, or a combination thereof.
- [00051] In another related embodiment, the invention is directed to a method for monitoring the effectiveness of treatment of a patient with an identified pancreatic lesion. The method comprises:
- [00052] (a) obtaining an initial cell-free DNA sample from a patient who is being treated;
- [00053] (b) enriching for hydroxymethylated DNA in the sample;
- [00054] (c) quantifying the nucleic acids in the enriched initial sample that map to each of a plurality of selected loci in a reference hydroxymethylation profile, wherein each selected locus comprises a hydroxymethylation biomarker;
- [00055] (d) comparing, at each locus, the hydroxymethylation level of the enriched cell-free DNA in the initial sample with the hydroxymethylation level in the reference profile, to ascertain differences in hydroxymethylation levels between the sample and the reference profile for each biomarker;
- [00056] (e) generating an initial hydroxymethylation profile for the patient comprising the hydroxymethylation level of the enriched cell-free DNA in the initial sample, at each locus;

[00057] (f) repeating steps (a) through (c) at a later time with a subsequent cell-free DNA sample obtained from the patient;

[00058] (g) generating a subsequent hydroxymethylation profile for the patient comprising the hydroxymethylation level of the enriched cell-free DNA in the subsequent sample, at each locus;

[00059] (h) comparing, at each locus, the hydroxymethylation level of the enriched cell-free DNA in the subsequent sample to the hydroxymethylation level of the enriched cell-free DNA in the initial sample, to ascertain a change in the pancreatic lesion; and

[00060] (i) if the comparison in step (e) evidences changes in the patient's hydroxymethylation profile that correlate with a progression toward cancer, changing the treatment protocol.

[00061] The progression toward cancer may involve a change in lesion size, grade, shape, nodal involvement, invasiveness, or two or more of any of the foregoing.

[00062] In another embodiment, the invention provides a method for reducing the risk of unnecessary pancreatic surgery, i.e., for reducing the risk that a pancreatic lesion surgically removed from a patient is benign. The method comprises, prior to surgery:

[00063] (a) obtaining a cell-free DNA sample from the patient;

[00064] (b) enriching for hydroxymethylated DNA in the sample;

[00065] (c) quantifying the nucleic acids in the enriched sample that map to each of a plurality of selected loci in a reference hydroxymethylation profile, wherein each selected locus comprises a hydroxymethylation biomarker;

[00066] (d) comparing, at each locus, the hydroxymethylation level of the sample with the hydroxymethylation level in the reference profile, to ascertain differences in hydroxymethylation levels between the sample and the reference profile for each biomarker; and

[00067] (e) calculating an index value representing the risk that the pancreatic lesion is cancerous from the comparison in step (d) combined with at least one additional parameter correlated with the risk that an individual has pancreatic cancer; and

[00068] (f) carrying out surgical resection of the pancreatic lesion only if the index value is greater than a value corresponding to a low risk of cancer.

[00069] In another embodiment, the invention provides a kit for carrying out any of the methods described herein in the analysis of a cell-free DNA sample obtained from a patient, where the kit comprises: at least one reagent for the determination of hydroxymethylation

level at each of a plurality of hydroxymethylation biomarker loci in a cell-free DNA sample; a solid support for capturing affinity-tagged 5hmC-containing cell-free DNA in the sample; and written instructions for the use of the at least one reagent and the solid support in carrying out the method.

[00070] In one aspect of the embodiment, the kit further includes instructions for accessing and using software designed to perform modeling and prediction.

[00071] In an additional embodiment, the kit comprises: a DNA β -glucosyl transferase; UDP glucose modified with a chemoselective group; a biotin moiety; a solid support having a surface functionalized with a biotin-binding protein; an adaptor comprising a molecular barcode; and written instructions for carrying out the method. As with the preceding embodiment, the kit may additionally include instructions for accessing and using software designed to perform modeling and prediction.

[00072] In a further embodiment, the invention provides a method for determining the likelihood that an individual at risk for developing pancreatic cancer has pancreatic cancer. The method comprises the following steps:

[00073] (a) obtaining a cell-free DNA sample from the patient;

[00074] (b) enriching for hydroxymethylated DNA in the sample;

[00075] (c) quantifying the nucleic acids in the enriched sample that map to each of a plurality of selected loci in a reference hydroxymethylation profile, wherein each selected locus comprises a hydroxymethylation biomarker;

[00076] (d) comparing, at each locus, the hydroxymethylation level of the sample with the hydroxymethylation level in the reference profile, to ascertain differences in hydroxymethylation levels between the sample and the reference profile for each biomarker; and

[00077] (e) calculating an index value representing the likelihood that the individual has pancreatic cancer from the comparison in step (d).

[00078] In one aspect, the method further includes, prior to step (a), identifying the individual as being at risk for developing pancreatic cancer from one or more parameters selected from: an identified pancreatic lesion; pancreatic inflammation; jaundice; age; weight; gender; ethnicity; family history; genetic mutations; diabetes; physical activity; diet; pro-inflammatory cytokine levels; and cigarette smoking.

[00079] In another embodiment, an improved multi-cancer test is provided that determines the likelihood that an individual has pancreatic cancer and at least one additional type of

cancer, wherein the improvement comprises determining the likelihood that the individual has pancreatic cancer by:

[00080] (a) obtaining a cell-free DNA sample from the individual;

[00081] (b) enriching for hydroxymethylated DNA in the sample;

[00082] (c) quantifying the nucleic acids in the enriched sample that map to each of a plurality of selected loci in a reference hydroxymethylation profile, wherein each selected locus comprises a hydroxymethylation biomarker;

[00083] (d) comparing, at each locus, the hydroxymethylation level of the sample with the hydroxymethylation level in the reference profile, to ascertain differences in hydroxymethylation levels between the sample and the reference profile for each biomarker; and

[00084] (e) calculating an index value representing the likelihood that the individual has pancreatic cancer from the comparison in step (d).

[00085] The test may further include eliminating false positives, false negatives, or both false positives and false negatives for the at least one additional type of cancer prior to (a).

[00086] The at least one additional type of cancer can be any type of cancer, including, without limitation, bladder cancer; cancers of the blood and bone marrow; brain cancer; breast cancer; cervical cancer; colorectal cancer; esophageal cancer; liver cancer; lung cancer; ovarian cancer; prostate cancer; renal cancer; skin cancer; testicular cancer; thyroid cancer; and uterine cancer.

[00087] In one aspect of this embodiment, the at least one additional type of cancer is selected from breast cancer, colorectal cancer, lung cancer, and prostate cancer.

BRIEF DESCRIPTION OF THE DRAWINGS

[00088] FIG. 1 schematically depicts the study cohorts employed in Example 1 herein. Cohorts: PDAC, n = 51, Non-cancer, n=41. Pooled non-cancer replicates were included across multiple 5hmC assay processing and sequencing batches.

[00089] FIG. 2 schematically depicts the sample processing workflow used in Example 1, including two alternating gender-divided flow cell constructs for detection of sample swaps.

[00090] FIG. 3 is a histogram showing the mean peak counts of 5hmC loci across distinct genomic regions, for the two cohorts, PDAC and non-cancer (identified in the figures as "PDAC" and "NC," respectively). It may be seen that non-coding features have a larger number of peaks.

[00091] FIG. 4 is a histogram providing the results of the enrichment analysis described in Example 1, with the Y-axis value equal to the mean of $\log_2$ (cancer/non-cancer). The histogram shows that gene-based features, SINEs, and Alus are enriched in 5hmC in both cancer and non-cancer cohorts, whereas intergenic regions, LINEs, and L1s are depleted of 5hmC peaks.

[00092] FIG. 5 provides box plots depicting statistically significant changes of 5hmC peaks in pancreatic cancer samples relative to non-cancer samples, in the promoter, LINE elements, exons, 3'UTR, and translation termination sites; here, the Y-axis value equal to $\log_2$ (cancer/non-cancer). Promoter and LINE elements were found to exhibit a depletion of 5-hydroxymethylcytosine (i.e., a decrease in hydroxymethylation) in the cancer (PDAC) samples relative to the non-cancer samples, while 5hmC enrichment was observed in exons, 3'UTR, and translation termination sites. In each box plot herein, the line within the box represents the median of the data, while the lower limit of the box represents the lower quartile and the upper limit of the box represents the upper quartile. Normally distributed data are portrayed as an aligned dot plot with error bars representing standard deviation from the mean. The calculated p-values are provided above each plot.

[00093] FIG. 6 provides box plots depicting statistically significant changes of 5hmC peaks in functional regions across pancreatic stages.

[00094] FIG. 7 provides box plots depicting 5hmC peak depletion in H3K4me3 and H3K27ac histone marks in the PDAC cohort (top panel) and ongoing H3K4me3 depletion observed in later stage disease (bottom panel).

[00095] FIG. 8A and FIG. 8B show 5hmC occupancy in the PANC- 1 cell line and normal pancreas histone map depicting variable occupancy in H3K4Me3 (FIG. 8A) with depletion at the center of the mark and complementary increase in 5hmC in H3K4Me1 (FIG. 8B). The results support a preferential increase in gene transcription in the PDAC cohort. The Y-axis values are equal to the normalized density of 5hmC counts in 10 bp windows. The dotted red lines = PDAC patient, one per line; the dotted blue lines = non-cancer patients, one per line; the solid red line = the average density of normalized 5hmC counts across all PDAC patients; and the solid blue line = the average density of normalized 5hmC counts across all non-cancer patients.

[00096] FIG. 9 is an MA plot showing all differentially represented genes and a heatmap showing 5hmC representation on the most significant genes.

[00097] FIG. 10 is a histogram showing the results of gene set enrichment analysis (GSEA) using differentially 5hmC-enriched genes. The blue bars represent the ratio of all pathways exhibiting reduced hydroxymethylation levels, and the orange bars represent the ratio of all pathways exhibiting higher hydroxymethylation levels, in PDAC samples relative to non-cancer samples. GSEA reveals that greater than 20% of KEGG pathways are both up-represented and down-represented in hydroxymethylation levels in PDAC versus non-cancer samples. Also, greater than 30% of immune pathways were found to be down-represented in PDAC versus non-cancer samples. In FIG. 10, "Hallmark" refers to the Hallmark gene sets in the MSigDB collections; "C2" refers to curated gene sets inclusive of the Biocarta, KEGG and Reactome databases; "C5.BP" refers to the "biological processes" subset of the Gene Ontology (GO) Consortium annotated gene sets; "C6" MSigDB oncogenic signature of cellular pathways that are often dis-regulated in cancer; "C7" (also referred to as "immuneSigDB") refers to the database of gene sets that represent cell types, states, and perturbations within the immune system.

[00098] FIG. 11 is a dot plot providing the results of PCA carried out using log [counts per million] on 13,180 genes with a statistically significant (FDR=0.05) increase or decrease in 5hmC in PDAC versus non-cancer samples. The dot plot exhibits visible partitioning of PDAC samples from non-cancer samples.

[00099] FIG. 12 is a PCA dot plot carried out using log [counts per million] on 320 genes, a subset of the 13,180 genes that exhibited a statistically significant (FDR= 0.05) increase or decrease in 5hmC in PDA versus non-cancer samples, where the data was filtered for increased PDAC representation as follows: (1) $(\log_2 [5\text{hmC-PDAC}/5\text{hmC-non-cancer}]) \geq 0.58$; and (2) $\log_2 [\text{average representation}] \geq 5$. The dot plot again shows good partitioning of PDAC samples from non-cancer samples despite an order of magnitude smaller gene set than used in the generation of FIG. 11.

[000100] FIG. 13 is a heatmap depicting the hierarchical clustering results obtained using the 320 genes selected for the PCA of FIG. 12 (the genes represent rows in the heatmap), to show how labeled samples (columns in the heat map) can be partitioned using log(CPM) 5mC counts. The heatmap shows near-perfect partitioning of the data, which in this case was that used by Stanford University in Song et al. (2017) *Cell Research* 27:1231-42 (sometimes referred to herein as the "Stanford data").

[000101] FIG. 14 is also a heatmap, prepared as explained above for FIG. 13, but using the data of Li et al. (2017) *Cell Research* 27:1243-1257 (sometimes referred to herein as the

"Chicago data"). In contrast to the almost perfect partitioning of the Stanford data, the Chicago data gave somewhat incomplete partitioning.

[000102] FIG. 15 and FIG. 16 provide the results of predictive modeling using two regularization models, Elastic Net and Lasso. FIG. 15 represents the training performed with 75% of the data, and FIG. 16 represents the test performed on the remaining 25% of the data, as described in Example 1 herein.

[000103] FIG. 17 gives the probability scores derived from each sample in the training data set using the Elastic Net and Lasso regularization methods. Probability scores near 1 are predicted cancer samples, while probability scores close to zero are non-cancer samples. The red line identified Q3 probability score of the non-cancer samples.

[000104] FIG. 18 represents the validation of the predictive models used with the Li et al. (2017) (Chicago) and Song et al. (2017) (Stanford) PDAC and non-cancer data sets.

[000105] FIG. 19 illustrates in graph form the hydroxymethylation levels ("5hmC occupancy") at loci associated with histone biomarkers H3K4me3, H3K4me1, and H3K27ac, and the similarity to an existing histone map from PANC-1 cell lines (LeRoy et al. (2013) *Epigenetics & Chromatin* 6:20).

[000106] FIG. 20 provides hydroxymethylation biomarker data obtained using the methods documented in Example 1 herein. The table of FIG. 20 identifies the genes by name and chromosome location and includes normalized values obtained with glmnet, glmnet2, glmnetF, and glmnet2F regularization methods; glmnetF and glmnet2F coefficients; mean and standard deviation; mean and standard deviation for the cancer cohort (identified as Mean-C and SD-C, respectively); mean and standard deviation for the non-cancer cohort (Mean-NC and SD-NC, respectively); Vote, computed as the sum of the glmnetF and glmnet2F normalized values for each gene; and the ratio of cancer-to-non-cancer (C/NC) means.

[000107] FIG. 21 provides a list of hydroxymethylation biomarkers suitable for use in conjunction with the present invention, by gene name, location, and glmnet value, identified using Study Group 2 in Example 2.

[000108] FIG. 22 is analogous to FIG. 20 but provides the biomarker data for the 41 genes of Table 4, *infra*, using Study Group 3 in Example 3.

DETAILED DESCRIPTION OF THE INVENTION

[000109] 1. Terminology and overview:

[000110] Unless defined otherwise, all technical and scientific terms used herein have the meaning commonly understood by one of ordinary skill in the art to which the invention pertains. Specific terminology of particular importance to the description of the present invention is defined below. Other relevant terminology is defined in International Patent Publication No. WO 2017/176630 to Quake et al. for "Noninvasive Diagnostics by Sequencing 5-Hydroxymethylated Cell-Free DNA." The aforementioned patent publication as well as all other patent documents and publications referred to herein are expressly incorporated by reference.

[000111] In this specification and the appended claims, the singular forms "a," "an" and "the" include plural referents unless the context clearly dictates otherwise. Thus, for example, "an adapter" refers not only to a single adapter but also to two or more adapters that may be the same or different, "a template molecule" refers to a single template molecule as well as a plurality of template molecules, and the like.

[000112] Numeric ranges are inclusive of the numbers defining the range. Unless otherwise indicated, nucleic acids are written left to right in 5' to 3' orientation; amino acid sequences are written left to right in amino to carboxy orientation, respectively.

[000113] The headings provided herein are not limitations of the various aspects or embodiments of the invention. Accordingly, the terms defined immediately below are more fully defined by reference to the specification as a whole.

[000114] The term "sample" as used herein relates to a material or mixture of materials, typically, although not necessarily, in liquid form, containing one or more analytes of interest.

[000115] The term "biological sample" as used herein relates to a sample derived from a biological fluid, cell, tissue, or organ of a human subject, comprising a mixture of biomolecules including proteins, peptides, lipids, nucleic acids, and the like. Generally, although not necessarily, the sample is a blood sample such as a whole blood sample, a serum sample, or a plasma sample, or a sample of pancreatic cyst fluid.

[000116] A "nucleic acid sample" as that term is used herein refers to a biological sample comprising nucleic acids. The nucleic acid sample may be a cell-free nucleic acid sample that comprises nucleosomes, in which case the nucleic acid sample is sometimes referred to herein as a "nucleosome sample." The nucleic acid sample may also be comprised of cell-free

DNA wherein the sample is substantially free of histones and other proteins, such as will be the case following cell-free DNA purification. The nucleic acid samples herein may also contain cell-free RNA.

[000117] A "sample fraction" refers to a subset of an original biological sample, and may be a compositionally identical portion of the biological sample, as when a blood sample is divided into identical fractions. Alternatively, the sample fraction may be compositionally different, as will be the case when, for example, certain components of the biological sample are removed, with extraction of cell-free nucleic acids being one such example.

[000118] As used herein, the term "cell-free nucleic acid" encompasses both cell-free DNA and cell-free RNA, where the cell-free DNA and cell-free RNA may be in a cell-free fraction of a biological sample comprising a body fluid. The body fluid may be blood, including whole blood, serum, or plasma, or it may be urine, cyst fluid, or another body fluid. In many instances, the biological sample is a blood sample, and a cell-free nucleic acid sample is extracted therefrom using now-conventional means known to those of ordinary skill in the art and/or described in the pertinent texts and literature; kits for carrying out cell-free nucleic acid extraction are commercially available (e.g., the AllPrep® DNA/RNA Mini Kit and QIAmp DNA Blood Mini Kit, both available from Qiagen, or the MagMAX Cell-Free Total Nucleic Acid Kit and the MagMAX DNA Isolation Kit, available from ThermoFisher Scientific). Also see, e.g., Hui et al. Fong et al. (2009) *Clin. Chem.* 55(3):587-598

[000119] The term "nucleotide" is intended to include those moieties that contain not only the known purine and pyrimidine bases, but also other heterocyclic bases that have been modified. Such modifications include methylated purines or pyrimidines, acylated purines or pyrimidines, alkylated riboses or other heterocycles. In addition, the term "nucleotide" includes those moieties that contain hapten or fluorescent labels and may contain not only conventional ribose and deoxyribose sugars, but other sugars as well. Modified nucleosides or nucleotides also include modifications on the sugar moiety, e.g., wherein one or more of the hydroxyl groups are replaced with halogen atoms or aliphatic groups, or are functionalized as ethers, amines, or the like. Of particular interest herein are modified cytosine residues, including 5-methylcytosine and oxidized forms thereof, such as 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxymethylcytosine.

[000120] The term "nucleic acid" and "polynucleotide" are used interchangeably herein to describe a polymer of any length, e.g., greater than about 2 bases, greater than about 10 bases, greater than about 100 bases, greater than about 500 bases, greater than 1000 bases, and up to

about 10,000 or more bases composed of nucleotides, e.g., deoxyribonucleotides or ribonucleotide. Nucleic acids may be produced enzymatically, chemically synthesized, or naturally obtained.

[000121] The term "oligonucleotide" as used herein denotes a single-stranded multimer of nucleotide of from about 2 to 200 nucleotides, up to 500 nucleotides in length.

[000122] Oligonucleotides may be synthetic or may be made enzymatically, and, in some embodiments, are 30 to 150 nucleotides in length. Oligonucleotides may contain ribonucleotide monomers (i.e., may be oligoribonucleotides) and/or deoxyribonucleotide monomers. An oligonucleotide may be 10 to 20, 21 to 30, 31 to 40, 41 to 50, 51 to 60, 61 to 70, 71 to 80, 80 to 100, 100 to 150 or 150 to 200 nucleotides in length, for example.

[000123] The term "hybridization" refers to the process by which a strand of nucleic acid joins with a complementary strand through base pairing as known in the art. A nucleic acid is considered to be "selectively hybridizable" to a reference nucleic acid sequence if the two sequences specifically hybridize to one another under moderate to high stringency hybridization and wash conditions. Moderate and high stringency hybridization conditions are known (see, e.g., Ausubel, et al., Short Protocols in Molecular Biology, 3rd ed., Wiley & Sons 1995 and Sambrook et al., Molecular Cloning: A Laboratory Manual, Third Edition, 2001 Cold Spring Harbor, N.Y.).

[000124] The terms "duplex" and "duplexed" are used interchangeably herein to describe two complementary polynucleotides that are base-paired, i.e., hybridized together. A DNA duplex is referred to herein as "double-stranded DNA" or "dsDNA" and may be an intact molecule or a molecular segment. For example, the dsDNA herein referred to as barcoded and adapter-ligated is an intact molecule, while the dsDNA formed between the nucleic acid tails of proximity probes in a proximity extension assay is a dsDNA segment.

[000125] The term "strand" as used herein refers to a single strand of a nucleic acid made up of nucleotides covalently linked together by covalent bonds, e.g., phosphodiester bonds. In a cell, DNA usually exists in a double-stranded form, and as such, has two complementary strands of nucleic acid referred to herein as the "top" and "bottom" strands. In certain cases, complementary strands of a chromosomal region may be referred to as "plus" and "minus" strands, "positive" and "negative" strands, the "first" and "second" strands, the "coding" and "noncoding" strands, the "Watson" and "Crick" strands or the "sense" and "antisense" strands. The assignment of a strand as being a top or bottom strand is arbitrary and does not imply any particular orientation, function or structure. The nucleotide sequences of the first

strand of several exemplary mammalian chromosomal regions (e.g., BACs, assemblies, chromosomes, etc.) is known, and may be found in NCBI's Genbank database, for example.

[000126] "Adapters" as that term is used herein are short synthetic oligonucleotides that serve a specific purpose in a biological analysis. Adapters can be single-stranded or double-stranded, although the preferred adapters herein are double-stranded. In one embodiment, an adapter may be a hairpin adapter (i.e., one molecule that base pairs with itself to form a structure that has a double-stranded stem and a loop, where the 3' and 5' ends of the molecule ligate to the 5' and 3' ends of a double-stranded DNA molecule, respectively). In another embodiment, an adapter may be a Y-adapter. In another embodiment, an adapter may itself be composed of two distinct oligonucleotide molecules that are base paired with each other. As would be apparent, a ligatable end of an adapter may be designed to be compatible with overhangs made by cleavage by a restriction enzyme, or it may have blunt ends or a 5' T overhang. The term "adapter" refers to double-stranded as well as single-stranded molecules. An adapter can be DNA or RNA, or a mixture of the two. An adapter containing RNA may be cleavable by RNase treatment or by alkaline hydrolysis. An adapter may be 15 to 100 bases, e.g., 50 to 70 bases, although adapters outside of this range are envisioned.

[000127] The term "adapter-ligated," as used herein, refers to a nucleic acid that has been ligated to an adapter. The adapter can be ligated to a 5' end and/or a 3' end of a nucleic acid molecule. As used herein, the term "adding adapter sequences" refers to the act of adding an adapter sequence to the end of fragments in a sample. This may be done by filling in the ends of the fragments using a polymerase, adding an A tail, and then ligating an adapter comprising a T overhang onto the A-tailed fragments. Adapters are usually ligated to a DNA duplex using a ligase, while with RNA, adapters are covalently or otherwise attached to at least one end of a cDNA duplex preferably in the absence of a ligase.

[000128] The term "asymmetric adapter", as used herein, refers to an adapter that, when ligated to both ends of a double stranded nucleic acid fragment, will lead to a top strand that contains a 5' tag sequence that is not the same as or complementary to the tag sequence at the 3' end. Examples of asymmetric adapters are described in U.S. Patents 5,712,126 and 6,372,434 to Weissman et al., and International Patent Publication No. WO 2009/032167 to Bignell et al. An asymmetrically tagged fragment can be amplified by two primers: a first primer that hybridizes to a first tag sequence added to the 3' end of a strand; and a second primer that hybridizes to the complement of a second tag sequence added to the 5' end of a

strand. Y-adapters and hairpin adapters (which can be cleaved, after ligation, to produce a "Y-adapter") are examples of asymmetric adapters.

[000129] The term "Y-adapter" refers to an adapter that contains: a double-stranded region and a single-stranded region in which the opposing sequences are not complementary. The end of the double-stranded region can be joined to target molecules such as double-stranded fragments of genomic DNA, e.g., by ligation or a transposase-catalyzed reaction. Each strand of an adapter-tagged double-stranded DNA that has been ligated to a Y-adapter is asymmetrically tagged in that it has the sequence of one strand of the Y-adapter at one end and the other strand of the Y-adapter at the other end. Amplification of nucleic acid molecules that have been joined to Y-adapters at both ends results in an asymmetrically tagged nucleic acid, i.e., a nucleic acid that has a 5' end containing one tag sequence and a 3' end that has another tag sequence.

[000130] The term "hairpin adapter" refers to an adapter that is in the form of a hairpin. In one embodiment, after ligation the hairpin loop can be cleaved to produce strands that have non-complementary tags on the ends. In some cases, the loop of a hairpin adapter may contain a uracil residue, and the loop can be cleaved using uracil DNA glycosylase and endonuclease VIII, although other methods are known.

[000131] The term "adapter-ligated sample", as used herein, refers to a sample that has been ligated to an adapter. As would be understood given the definitions above, a sample that has been ligated to an asymmetric adapter contains strands that have non-complementary sequences at the 5' and 3' ends.

[000132] The term "amplifying" as used herein refers to generating one or more copies, or "amplicons," of a template nucleic acid, such as may be carried out using any suitable nucleic acid amplification technique, such as technology, such as PCR, NASBA, TMA, and SDA.

[000133] The terms "enrich" and "enrichment" refer to a partial purification of template molecules that have a certain feature (e.g., nucleic acids that contain 5-hydroxymethylcytosine) from analytes that do not have the feature (e.g., nucleic acids that do not contain hydroxymethylcytosine). Enrichment typically increases the concentration of the analytes that have the feature by at least 2-fold, at least 5-fold or at least 10-fold relative to the analytes that do not have the feature. After enrichment, at least 10%, at least 20%, at least 50%, at least 80% or at least 90% of the analytes in a sample may have the feature used for enrichment. For example, at least 10%, at least 20%, at least 50%, at least 80% or at least

90% of the nucleic acid molecules in an enriched composition may contain a strand having one or more hydroxymethylcytosines that have been modified to contain a capture tag.

[000134] The term "sequencing," as used herein, refers to a method by which the identity of at least 10 consecutive nucleotides (e.g., the identity of at least 20, at least 50, at least 100 or at least 200 or more consecutive nucleotides) of a polynucleotide is obtained.

[000135] The terms "next-generation sequencing" (NGS) or "high-throughput sequencing", as used herein, refer to the so-called parallelized sequencing-by-synthesis or sequencing-by-ligation platforms currently employed by Illumina, Life Technologies, Roche, etc. Next-generation sequencing methods may also include nanopore sequencing methods such as that commercialized by Oxford Nanopore Technologies, electronic detection methods such as Ion Torrent technology commercialized by Life Technologies, and single-molecule fluorescence-based methods such as that commercialized by Pacific Biosciences.

[000136] The term "read" as used herein refers to the raw or processed output of sequencing systems, such as massively parallel sequencing. In some embodiments, the output of the methods described herein is reads. In some embodiments, these reads may need to be trimmed, filtered, and aligned, resulting in raw reads, trimmed reads, aligned reads.

[000137] A "Unique Feature Identifier" (UFI) sequence refers to a relatively short nucleic acid sequence that serves to identify a feature of a nucleic acid molecule. Nucleic acid template molecules and amplicons thereof that contain a UFI are sometimes referred to herein as "barcoded" template molecules or amplicons. Examples of UFI sequence types include, without limitation, the following:

[000138] A "source identifier sequence" (or "source UFI" or "source barcode") identifies the biological sample (or other source) of origin. That is, each DNA molecule in a single sample is tagged with the same source identifier sequence, thus allowing the mixing of samples prior to sequencing. These UFIs may also be characterized as a "sample identifier sequence," a "sample UFI," or "sample barcode."

[000139] A "fragment identifier sequence" (or "fragment UFI" or "fragment barcode"): In a nucleic acid sample in which nucleic acids have been fragmented, each fragment in a sample is barcoded with a corresponding fragment identifier sequence. Sequence reads that have non-overlapping fragment identifier sequences represent different original nucleic acid template molecules, while reads that have the same fragment identifier sequences, or substantially overlapping fragment identifier sequences, likely represent fragments of the

same template molecule. The unique feature identified here is the template nucleic acid molecule from which a fragment derives.

[000140] A "strand identifier sequence" (or "strand UFI" or "strand barcode") independently tags each of the two strands of a DNA duplex, so that the strand from which a read originates can be determined, i.e., as the W strand or the C strand.

[000141] A "5hmC identifier sequence" (or "5hmC barcode") identifies DNA fragments originating from 5hmC-containing cell-free DNA template molecules in a sample, i.e., "hydroxymethylated" DNA.

[000142] A "5mC identifier sequence" (or "5mC barcode") identifies DNA fragments originating from 5mC-containing cell-free DNA template molecules that do not contain 5hmC.

[000143] A "molecular UFI sequence" (or "molecular barcode") is appended to every nucleic acid template molecule in a sample, and is random, such that, providing the UFI sequence is of sufficient length, every nucleic acid template molecule is attached to a unique UFI sequence. Molecular UFI sequences, as is known in the art, can be used to account for and offset amplification and sequencer errors, allow a user to track duplicates and remove them from downstream analysis, and enable molecular counting, and, in turn, the determination of an analyte concentration. See, e.g., Casbon et al. (2011) *Nuc. Acids Res.* 39(12):1-8. The "unique feature" here is the identity of the nucleic acid template molecules.

[000144] In some embodiments, a UFI may have a length in the range of from 1 to about 35 nucleotides, e.g., from 3 to 30 nucleotides, 4 to 25 nucleotides, or 6 to 20 nucleotides. In certain cases, the UFI may be error-detecting and/or error-correcting, meaning that even if there is an error (e.g., if the sequence of the molecular barcode is mis-synthesized, mis-read or distorted during any of the various processing steps leading up to the determination of the molecular barcode sequence) then the code can still be interpreted correctly. The use of error-correcting sequences is described in the literature (e.g., in U.S. Patent Publication Nos. U.S. 2010/0323348 to Hamati et al. and U.S. 2009/0105959 to Braverman et al., both of which are incorporated herein by reference).

[000145] The oligonucleotides that serve as UFI sequences herein may be incorporated into DNA molecule using any effective means, where "incorporated into" is used interchangeably herein with "added to" and "appended to," insofar as the UFI can be provided at the end of a DNA molecule, near the end of a DNA molecule, or within a DNA molecule. For example, multiple UFIs can be end-ligated to DNA using a selected ligase, in which case only the final

UFI is at the "end" of the molecule. In addition, in the proximity extension assay and histone modification methods described in detail *infra*, the UFI may be contained within the nucleic acid tail of a proximity probe, at the end of the nucleic acid tail of a proximity probe, or within the hybridized region generated upon the binding of probes to the protein target.

[000146] More generally, the term "detection" is used interchangeably with the terms "determining," "measuring," "evaluating," "assessing," "assaying," and "analyzing," to refer to any form of measurement, and include determining if an element is present or not. These terms include both quantitative and/or qualitative determinations. Assessing may be relative or absolute. "Assessing the presence of" thus includes determining the amount of a moiety present, as well as determining whether it is present or absent. Assessing the level at a hydroxymethylation biomarker locus refers to a determination of the degree of hydroxymethylation at that locus.

[000147] "Accuracy" refers to the degree of conformity of a measured or calculated quantity (a test reported value) to its accurate (or true) value. Clinical accuracy relates to the proportion of true outcomes (true positives (TP) or true negatives (TN) versus misclassified outcomes (false positives (FP) or false negatives (FN), and may be stated as a sensitivity, specificity, positive predictive values (PPV) or negative predictive values (NPV), or as a likelihood, or odds ratio, among other measures.

[000148] "Performance" is a term that relates to the overall usefulness and quality of a diagnostic or prognostic test, including, among others, clinical and analytical accuracy, other analytical and process characteristics, such as use characteristics (e.g., stability, ease of use), health economic value, and relative costs of components of the test. Any of these factors may be the source of superior performance and thus usefulness of the test, and may be measured by appropriate "performance metrics," such as AUC, time to result, shelf life, etc. as relevant.

[000149] "Clinical parameters" encompass all non-sample biomarkers of subject health status or other characteristics, such as, without limitation, lesion size; lesion location; presence or absence of pancreatic inflammation; presence or absence of other symptoms; patient age; weight; jaundice; gender; ethnicity; family history; genetic mutations; diabetes mellitus (including Type I and Type II diabetes); physical activity; diet; pro-inflammatory cytokine levels; and smoking status of the patient.

[000150] A "formula," "algorithm," or "model" is any mathematical equation, algorithmic, analytical or programmed process, or statistical technique that takes one or more continuous or categorical inputs and calculates an output value, sometimes referred to as an "index" or

"index value." Non-limiting examples of "formulas" include sums, ratios, and regression operators, such as coefficients or exponents, biomarker value transformations and normalizations (including, without limitation, those normalization schemes based on clinical parameters, such as gender, age, or ethnicity), rules and guidelines, statistical classification models, and neural networks trained on historical populations. Of particular use in combining hydroxymethylation levels at various biomarker loci and clinical parameters, optionally in further combination with other factors (e.g., non-hydroxymethylation biomarkers), are linear and non-linear equations and statistical classification analyses to determine the relationship between hydroxymethylation levels at the biomarker loci detected in a patient sample and the patient's risk of having or developing pancreatic cancer. In panel and combination construction, of particular interest are structural and syntactic statistical classification algorithms, and methods of risk index construction, utilizing pattern recognition and machine learning features, including established techniques such as cross-correlation, Principal Components Analysis (PCA), factor rotation, Logistic Regression (LogReg), Linear Discriminant Analysis (LDA), Eigengene Linear Discriminant Analysis (ELDA), Support Vector Machines (SVM), Random Forest (RF), Recursive Partitioning Tree (RPART), as well as other related decision tree classification techniques, Shrunk Centroids (SC), StepAIC, Kth-Nearest Neighbor, Boosting, Decision Trees, Neural Networks, Bayesian Networks, Support Vector Machines, and Hidden Markov Models, among others. Many such algorithmic techniques have been further implemented to perform both feature (loci) selection and regularization, such as in ridge regression, lasso, and elastic net, among others. Other techniques may be used in survival and time to event hazard analysis, including Cox, Weibull, Kaplan-Meier and Greenwood models well known to those of skill in the art. Many of these techniques are useful either combined with a hydroxymethylation biomarker selection technique, such as forward selection, backwards selection, or stepwise selection, complete enumeration of all potential biomarker sets, or panels, of a given size, genetic algorithms, or they may themselves include biomarker selection methodologies. These may be coupled with information criteria, such as Akaike's Information Criterion (AIC) or Bayes Information Criterion (BIC), in order to quantify the tradeoff between additional biomarkers and model improvement, and to aid in minimizing overfit. The resulting predictive models may be validated in other studies, or cross-validated in the study they were originally trained in, using such techniques as Bootstrap, Leave-One-Out (LOO) and 10-Fold cross-validation

(10-Fold CV). At various steps, false discovery rates may be estimated by value permutation according to techniques known in the art.

[000151] "Risk," in the context of the present invention, relates to the probability that an event will occur over a specific time period, as in the development of pancreatic cancer, and can mean a subject's "absolute" risk or "relative" risk. Absolute risk can be measured with reference to index values developed from statistically valid historical cohorts that have been followed for the relevant time period; an example of absolute risk herein is knowledge of the outcome of a pancreatic biopsy following surgical resection. Relative risk refers to the ratio of absolute risks of a subject compared either to the absolute risks of low risk cohorts.

[000152] "Risk evaluation" or "evaluation of risk" in the context of the present invention encompasses making a prediction of the probability, odds, or likelihood that an event or disease state may occur, the rate of occurrence of the event or conversion from one state to another, i.e., from an apparently benign pancreatic lesion to a cancerous lesion, and the like. The methods of the present invention may be used to make continuous or categorical measurements of the risk of conversion of an apparently benign pancreatic lesion to a cancerous lesion. In the categorical scenario, the invention can be used to discriminate between normal and other subject cohorts at higher risk for developing pancreatic cancer. In other embodiments, the present invention may be used so as to discriminate those at risk for developing pancreatic cancer from those having pancreatic cancer, or those likely to respond well to a particular treatment from those who are not. Such differing uses may require different hydroxymethylation biomarker combinations and individualized panels, mathematical algorithms, and/or cut-off points, but be subject to the same measurements of accuracy and performance for the respective intended use.

[000153] A "hydroxymethylation level" or "hydroxymethylation state" is the extent of hydroxymethylation within a hydroxymethylation biomarker locus. The extent of hydroxymethylation is normally measured as hydroxymethylation density, e.g., the ratio of 5hmC residues to total cytosines, both modified and unmodified, within a nucleic acid region. Other measures of hydroxymethylation density are also possible, e.g., the ratio of 5hmC residues to total nucleotides in a nucleic acid region.

[000154] A "hydroxymethylation profile" or "hydroxymethylation signature" refers to a data set that comprises the hydroxymethylation level at each of a plurality of hydroxymethylation biomarker loci. The hydroxymethylation profile may be a reference hydroxymethylation profile that comprises composite a hydroxymethylation profile for a

population of individuals with at least one shared characteristic, as explained *infra*. The hydroxymethylation profile may also be a patient hydroxymethylation profile, constructed from the measurement of hydroxymethylation levels at each of a plurality of hydroxymethylation biomarker sites.

[000155] A "reference hydroxymethylation profile" thus refers to a data set representing the hydroxymethylation level of each of a plurality of hydroxymethylation biomarkers, where the data set is a composite of hydroxymethylation profiles of a plurality of individuals having at least one shared characteristic, e.g., individuals who have had a pancreatic lesion identified in an imaging scan, individuals who have not had a pancreatic lesion identified in an imaging scan, individuals who have not had pancreatic cancer, individuals with chronic pancreatitis, and the like.

[000156] The "hydroxymethylation biomarkers" herein comprise loci selected for their relevance to pancreatic cancer, particularly an exocrine pancreatic cancer such as PDAC. By "relevance" is meant that a hydroxymethylation biomarker locus, alone or in combination with one or more other hydroxymethylation biomarker loci, tends to exhibit an increase or decrease in hydroxymethylation in a manner that correlates with the risk, presence, absence, type, size, stage, invasiveness, grade, location, diagnosis, prognosis, outcome, and/or or likelihood of treatment responsiveness of pancreatic cancer, including the determinations of any of steps (a) through (j) in the preceding section.

[000157] The term "locus" in the preceding paragraph and throughout this application refers to a site on a nucleic acid molecule, wherein the nucleic acid molecule may be single-stranded or double-stranded, and further wherein an individual locus (or multiple "loci") may be of any length, thus including a single CpG site as well as a full-length gene, or across larger features such as topologically associated domains, including when several such loci are aggregated into groups such as related sequence motifs, other homologies or functional characteristics (regardless of their adjacency or topological relationship). The loci herein may be contained within a gene body; within an annotation feature outside of the gene body, such as a promoter, an enhancer, a transcription initiation site, a transcription stop site, or a DNA binding site, or a combination thereof; or within an untranslated region, or "UTR" (including 3'UTRs and 5'UTRs). DNA binding sites that may contain one or more reference loci include, by way of example, silenced regions, transcription factor binding sites, transcription repressor binding sites, and CTCF binding sites (transposon repeat regions). Reference loci within CTCF binding sites are of particular interest, insofar as the CTCF gene codes for

transcriptional repressor CTCF (also known as 11-zinc finger protein or CCCTC-binding factor), which in turn is involved in many cellular processes, including transcription regulation and regulation of chromatin architecture. See, for example, Juan et al. (2016) *Cell Reports* 14(5): 1246-1257; and Escedi et al. (2018) *Epigenomes* 2(1):3.

[000158] It should be noted that some of the individual hydroxymethylation biomarkers disclosed herein may not have significant individual significance in the evaluation of a pancreatic lesion, but when used in combination with other hydroxymethylation biomarkers disclosed herein and clinical parameters impacting on the evaluation and monitoring of a pancreatic lesion, optionally further combined with one or more other types of biomarkers and/or patient-specific risk factors, become significant in discriminating as a method of the invention requires, e.g., between a subject who has pancreatic cancer and a subject who does not have pancreatic cancer, or between a subject who is likely to develop pancreatic cancer and a subject who is not likely to develop pancreatic cancer, etc. The methods of the present invention provide an improvement over currently available methods of evaluating the risk that a subject has pancreatic cancer or is likely to develop pancreatic cancer, by using the biomarkers defined herein. To the extent that other biomarker pathway participants (i.e., other biomarker participants in common pathways with those biomarkers contained within the list of hydroxymethylation biomarkers herein are also relevant pathway participants in the subject pancreatic conditions, they may be functional equivalents to the hydroxymethylation biomarkers thus far disclosed. Furthermore, other unlisted hydroxymethylation biomarkers will be very highly correlated with the individual hydroxymethylation biomarkers listed here (for the purpose of this application, any two variables will be considered to be “very highly correlated” when they have a Coefficient of Determination (R^2) of 0.5 or greater). The present invention encompasses such functional and statistical equivalents to the aforementioned hydroxymethylation biomarkers. Furthermore, the statistical utility of such additional hydroxymethylation biomarkers is substantially dependent on the cross-correlation between multiple biomarkers and any new biomarkers will often be required to operate within a panel in order to elaborate the meaning of the underlying biology.

[000159] The term “correlate” as used herein in reference to a variable (e.g., a value, a set of values, a disease state, a risk associated with the disease state, or the like) is a measure of the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel. One example of a positive correlation is the relationship between a hydroxymethylation level at a hydroxymethylation

biomarker locus, on the one hand, and the risk of developing pancreatic cancer, on the other, when the hydroxymethylation level increases as the risk of developing cancer increases. Conversely, a negative correlation would exist when the hydroxymethylation level biomarker at a hydroxymethylation biomarker locus decreases as the risk of developing cancer increases.

[000160] The term "pancreatic cancer" herein refers to an exocrine pancreatic cancer, particularly PDAC.

[000161] The present invention relates, in part, to the discovery that certain biological markers, particularly epigenetic markers relating to DNA hydroxymethylation, correlate in some way with pancreatic cancer, particularly an exocrine cancer such as PDAC. The methods involve measuring the hydroxymethylation level at each of a plurality of hydroxymethylation biomarker loci to generate a hydroxymethylation profile for a patient, and then comparing the patient's hydroxymethylation profile to a reference hydroxymethylation profile, at each locus. The biomarkers are differentially hydroxymethylated in subjects who have pancreatic cancer or are at risk of developing pancreatic cancer, particularly PDAC or another exocrine pancreatic cancer.

[000162] In some embodiments, the invention enables the determination of the risk that a pancreatic lesion observed with an imaging scan, i.e., an identified pancreatic lesion, is cancerous; the risk that an identified noncancerous pancreatic lesion will become cancerous; the likelihood that a particular therapy for treating a subject with pancreatic cancer will be effective; the risk that a subject without an identified pancreatic lesion will, at some point, develop a pancreatic lesion, as well as the risk of that lesion becoming cancerous.

[000163] The invention also enables a practitioner to determine the effectiveness of a therapy a subject is undergoing in connection with an identified pancreatic lesion; an increase or decrease in the risk that an identified pancreatic lesion will develop into cancer; an increase or decrease in the likelihood that a subject without an observed pancreatic lesion will develop a pancreatic lesion, and the risk of that lesion becoming cancerous; and a change in an identified pancreatic lesion, including a change in the size, stage, grade, or degree of invasiveness of a cancerous pancreatic lesion.

[000164] *2. Determination of hydroxymethylation profile:*

[000165] Each embodiment of the invention comprises, initially, the generation of a patient's hydroxymethylation profile. The profile is generated by ascertaining the hydroxymethylation level at each of a plurality of hydroxymethylation biomarker loci, and

assembling the data so obtained into a data set that serves as the hydroxymethylation profile. The hydroxymethylation biomarkers are differentially hydroxymethylated in subjects who have pancreatic cancer or who are at risk of developing pancreatic cancer, relative to a reference hydroxymethylation profile. That is, the biomarkers comprise regions of genomic DNA that are more susceptible to increases or decreases in hydroxymethylation level than other regions of the DNA, and that exhibit an increase or decrease in hydroxymethylation level in a manner that correlates with pancreatic cancer or a risk of developing pancreatic cancer.

[000166] In a first embodiment, the invention provides a method for evaluating the risk that a pancreatic lesion identified on an imaging scan is cancerous. The imaging may be carried out using any suitable method, although cross-sectional imaging is preferred, e.g., using multi-detector row computed tomography (CT) or magnetic resonance imaging (MRI) with MR cholangiopancreatography (MRCP).

[000167] The first step in the method involves obtaining a cell-free DNA (cfDNA) sample from a blood sample or cyst fluid sample taken from the patient. Extraction of cfDNA can be carried out using any suitable technique, for example using the commercially available kits referenced in the preceding section. The cfDNA is then enriched, so that the concentration of the cfDNA is substantially increased, a virtual necessity because of the very low levels of cfDNA normally obtained. A generally preferred enrichment technique is described in International Patent Publication WO 2017/176630 to Quake et al., incorporated herein by reference in its entirety: an affinity tag is appended to 5hmC residues in a sample of cfDNA, and the tagged DNA molecules are then selectively removed by bonding to a functionalized solid support. An illustrative example of the method, as described in Quake et al., involves initially modifying end-blunted, adaptor-ligated double-stranded DNA fragments in the cell-free sample to covalently attach biotin, as the affinity tag, to 5hmC residues. This may be carried out by selectively glucosylating 5hmC residues with uridine diphospho (UDP) glucose functionalized at the 6-position with an azide moiety, a step that is followed by a spontaneous 1,3-cycloaddition reaction with alkyne-functionalized biotin via a "click chemistry" reaction. The DNA fragments containing the biotinylated 5hmC residues are adapter-ligated dsDNA template molecules that can then be pulled down with a solid support functionalized with a biotin-binding protein (e.g., avidin or streptavidin) in the enrichment step.

[000168] The cfDNA is then amplified without releasing the captured cfDNA from the support, thereby giving a plurality of amplicons. Any suitable amplification technique may be employed (e.g., PCR, NASBA, TMA, SDA) although PCR is preferred.

[000169] Next, the nucleic acids that map to each of a plurality of selected loci in a reference hydroxymethylation profile are quantified, so that after amplification, pooling, and sequencing, information regarding hydroxymethylation levels can be deduced from the sequence reads obtained. That is, the sequence reads are analyzed to provide a quantitative determination of which sequences are hydroxymethylated in the cfDNA, and the level of hydroxymethylation. This may be done by, e.g., counting sequence reads or, alternatively, counting the number of original starting molecules, prior to amplification, based on their fragmentation breakpoint and/or whether they contain the same molecular UFI. The use of molecular UFI sequences (or "molecular barcodes" as they are sometimes called) in conjunction with other features of the fragments (e.g., the end sequences of the fragments, which define the breakpoints) to distinguish between the fragments is known. See Casbon (2011) *Nucl. Acids Res.* 22 e81 and Fu et al. (2011) *Proc. Natl. Acad. Sci. USA* 108: 9026-31), among others. Molecular barcodes are also described in U.S. Patent Publication Nos. 2015/0044687, 2015/0024950, and 2014/0227705, and in U.S. Patent Nos. 8,835,358 and US 7,537,897, as well as a variety of other publications.

[000170] The molecular UFI sequence is preferably incorporated into the adapters that are end-ligated to the cfDNA following extraction thereof. The adapters may be constructed so as to comprise an additional UFI sequence, e.g., a sample UFI sequence, a strand-identifier UFI sequence, or both.

[000171] Other methods of ascertaining the hydroxymethylation profile of DNA in the cell-free nucleic sample are described in Provisional U.S. Patent Application Serial No. 62/630,798 to Arensdorf for "Methods for the Epigenetic Analysis of DNA, particularly Cell-Free DNA," filed February 14, 2018, and in U.S. Patent Publication No. 2017/0298422 to Song et al., both of which are incorporated by reference herein. These references are also useful in conjunction with an embodiment of the invention in which the present combined workflow process further includes the detection of a cfDNA methylation profile in addition to the cfDNA hydroxymethylation profile.

[000172] The selected loci in the above described method are hydroxymethylation biomarkers, i.e., loci that have been identified herein as differentially hydroxymethylated in a manner that relates to the presence, absence, or risk of pancreatic cancer. Certain

hydroxymethylation biomarkers that are particularly useful in conjunction with the present methods, as established in Example 1, include, without limitation, those set forth in Table 1 (along with chromosome location):

[000173] *Table 1:*

Gene	Location
ADARB2-AS1	chr10
ANKRD36B	chr02
ASAH2B	chr10
ATG4B	chr02
ATP8B1	chr18
BOLA1	chr01
C11orf88	chr11
C17orf97	chr17
C1orf170	chr01
C3orf36	#N/A
C8orf74	chr08
CAMSAP2	chr01
CCDC54	chr03
CCDC59	chr12
CKAP2	chr13
CLK2P	#N/A
CRTC1	chr19
CSRP2	chr12
CYB5D1	chr17
DNAJC27	#N/A
DYNAP	#N/A
FAM166A	#N/A
FAM188B	chr07
FAM196A	chr10
FAM86JP	#N/A
FAT4	chr04
FBXO5	chr06
FGF2	chr04
FUT2	chr19
GAS2L2	chr17
GAS6	#N/A
GGACT	chr13
GLRX5	chr14
GPX1	#N/A
GPX5	chr06
HBD	chr11

Gene	Location
HLA-A	chr06
HTR1F	chr03
IL36G	#N/A
KANSL1	#N/A
KCNH6	#N/A
KCTD15	#N/A
KLHL38	#N/A
KLK2	#N/A
KRT6B	chr12
LAMC1	chr01
LGALS14	chr19
LGALS8-AS1	#N/A
LIFR	chr05
LINC00266-1	#N/A
LINC00310	chr21
LOC100130452	chr02
LOC100130557	#N/A
LOC100130894	#N/A
LOC100288778	#N/A
LOC100505633	#N/A
LOC100505648	chr15
LOC100505738	#N/A
LOC100652909	chr19
LOC389033	#N/A
LOC90784	#N/A
LRRC37A2	#N/A
MED11	chr17
MRPL23-AS1	chr11
NAT8L	chr04
NEUROD1	#N/A
NEUROG2	chr04
NME5	chr05
NOMO3	chr16
NPRL2	chr03
NXN	chr17
ODF3L1	#N/A
ODF3L2	chr19
OSCP1	chr01
PARD6G	chr18
PGAM1	#N/A
PLA2G2E	#N/A

Gene	Location
PLSCR4	#N/A
PPAP2A	chr05
PPP1R15A	chr19
PPP1R3E	#N/A
RASL10B	chr17
REXO1L1	#N/A
RIMBP3	#N/A
RNF126P1	chr17
RNU6-76	chr16
RPP25	chr15
RPS27	#N/A
SH3PXD2B	#N/A
SHISA4	#N/A
SLC25A38	#N/A
SLC4A1	#N/A
SLCO5A1	#N/A
SPDEF	chr06
SRSF6	#N/A
STRA6	#N/A
SYNM	chr15
TBCB	chr19
TDRD6	#N/A
TEX26	#N/A
TMEM253	#N/A
TNFSF13B	#N/A
TTC14	#N/A
TUBA4A	#N/A
UBB	chr17
VAMP8	chr02
VGLL2	#N/A
WASH2P	#N/A
WNT9B	#N/A
XBP1	chr22
ZNF789	#N/A

[000174] While the experimental work documented in Example 1 identified thousands of genes in which 5hmC is differentially expressed, the above group represents a stringently filtered set of the most significant genes using Elastic Net regularization (glmnetF) or Lasso regularization (glmnet2F). The above 111 genes were found to exhibit biology related to pancreatic development (GATA4, GATA6, PROX1, and ONECUT1) and/or cancer

development (YAP, TEAD1, PROX, ONECUT1, ONECUT2, IGF1, and IGF2), as explained in Example 1 herein. Table 2 indicates those genes identified using glmnetF and Table 3 indicates those genes identified using glmnet2F:

[000175] Table 2:

Gene	Location
ADARB2-AS1	chr10
ANKRD36B	chr02
ATG4B	chr02
ATP8B1	chr18
BOLA1	chr01
C11orf88	chr11
C17orf97	chr17
C1orf170	chr01
C3orf36	#N/A
C8orf74	chr08
CAMSAP2	chr01
CCDC54	chr03
CCDC59	chr12
CKAP2	chr13
CLK2P	#N/A
CRTC1	chr19
CSRP2	chr12
CYB5D1	chr17
DNAJC27	#N/A
DYNAP	#N/A
FAM166A	#N/A
FAM188B	chr07
FAM196A	chr10
FAM86JP	#N/A
FAT4	chr04
FBXO5	chr06
FGF2	chr04
FUT2	chr19
GAS2L2	chr17
GAS6	#N/A
GGACT	chr13
GLRX5	chr14
GPX1	#N/A
GPX5	chr06
HBD	chr11
HLA-A	chr06

Gene	Location
HTR1F	chr03
IL36G	#N/A
KANSL1	#N/A
KCNH6	#N/A
KCTD15	#N/A
KLHL38	#N/A
KLK2	#N/A
KRT6B	chr12
LAMC1	chr01
LGALS14	chr19
LGALS8-AS1	#N/A
LIFR	chr05
LINC00266-1	#N/A
LINC00310	chr21
LOC100130557	#N/A
LOC100130894	#N/A
LOC100288778	#N/A
LOC100505633	#N/A
LOC100505648	chr15
LOC100505738	#N/A
LOC100652909	chr19
LOC389033	#N/A
LOC90784	#N/A
LRRC37A2	#N/A
MED11	chr17
MRPL23-AS1	chr11
NAT8L	chr04
NEUROD1	#N/A
NEUROG2	chr04
NME5	chr05
NOMO3	chr16
NPRL2	chr03
NXN	chr17
ODF3L1	#N/A
ODF3L2	chr19
OSCP1	chr01
PARD6G	chr18
PGAM1	#N/A
PLA2G2E	#N/A
PLSCR4	#N/A
PPAP2A	chr05

Gene	Location
PPP1R15A	chr19
PPP1R3E	#N/A
RASL10B	chr17
REXO1L1	#N/A
RIMBP3	#N/A
RNF126P1	chr17
RNU6-76	chr16
RPP25	chr15
RPS27	#N/A
SH3PXD2B	#N/A
SHISA4	#N/A
SLC25A38	#N/A
SLC4A1	#N/A
SLCO5A1	#N/A
SPDEF	chr06
SRSF6	#N/A
STRA6	#N/A
SYNM	chr15
TBCB	chr19
TDRD6	#N/A
TEX26	#N/A
TMEM253	#N/A
TNFSF13B	#N/A
TTC14	#N/A
TUBA4A	#N/A
UBB	chr17
VAMP8	chr02
VGLL2	#N/A
WASH2P	#N/A
WNT9B	#N/A
XPB1	chr22
ZNF789	#N/A

[000176] Table 3:

Gene	Location
ASAH2B	chr10
BOLA1	chr01
C11orf88	chr11
C17orf97	chr17
C3orf36	#N/A

Gene	Location
CCDC54	chr03
CKAP2	chr13
CLK2P	#N/A
DNAJC27	chr02
DYNAP	#N/A
FAM166A	chr09
FGF2	chr04
FUT2	chr19
GAS6	#N/A
GGACT	chr13
GPX1	#N/A
GPX5	chr06
IL36G	#N/A
KCNH6	chr17
KCTD15	chr19
KLK2	#N/A
KRT6B	chr12
LGALS14	chr19
LGALS8-AS1	#N/A
LIFR	chr05
LINC00266-1	#N/A
LINC00310	chr21
LOC100130452	chr02
LOC100130557	#N/A
LOC100288778	#N/A
LOC100505648	chr15
MED11	chr17
NAT8L	chr04
NEUROG2	chr04
NME5	chr05
NOMO3	chr16
NPRL2	chr03
NXN	chr17
ODF3L1	#N/A
PPP1R3E	#N/A
RASL10B	chr17
RNF126P1	chr17
SH3PXD2B	chr05
SLC25A38	chr03
TNFSF13B	chr13
VAMP8	chr02

Gene	Location
VGLL2	#N/A

[000177] Other hydroxymethylation biomarkers that are useful in conjunction with the present methods are the 611 genes set forth in FIG. 21, along with location and glmnet value (identified using Study Group 2 in Example 2). Hydroxymethylation biomarkers within this group that may be of particular interest are the 41 biomarkers set forth in Table 4, again along with location and glmnet value (from Study Group 3 in Example 3):

[000178] *Table 4:*

Gene	Location	Glmnet
LINC00457	chr13~~~35009590~35214822~	100
CERS3	chr15~~~100940599~101084925~	67
LOC285629	chr05~~~160358785~160365633~	65
RHOJ	chr14~+~63671101~63760230~	57
GP2	chr16~~~20321810~20338835~	56
SFRP1	chr08~~~41119475~41166990~	56
LY6G6F	chr06~+~31674683~31678372~	51
HOXA4	chr07~~~27168125~27170399~	50
MYOCD	chr17~+~12569206~12670651~	46
C14orf64	chr14~~~98391946~98444461~	45
PDE10A	chr06~~~165740775~166075588~	42
PTCRA	chr06~+~42883726~42893575~	42
UCP3	chr11~~~73711325~73720282~	41
NTRK2	chr09~+~87283465~87638505~	36
GABRG3	chr15~+~27216428~27778373~	33
FBXL7	chr05~+~15500304~15939900~	31
LOC100128714	chr15~+~26147506~26298267~	23
LOC151171	chr02~~~239419330~239464140~	23
MIR5009	chr21~~~28659821~29283529~	20
HKR1	chr19~+~37825579~37855357~	17
ZNF573	chr19~~~38229202~38270230~	16
LINC00670	chr17~+~12453284~12540504~	16
FILIP1	chr06~~~76017799~76203496~	16
DCLK1	chr13~~~36342788~36705514~	15
OSGEP	chr14~~~20915206~20923267~	13
BNC2	chr09~~~16409500~16870786~	12
ABCC12	chr16~~~48116883~48180681~	11
PCDH7	chr04~+~30722029~31148423~	10
ZNF2	chr02~+~95831182~95850064~	10

Gene	Location	Glmnet
PDLIM1	chr10~~~96997324~97050905~	9
TSPAN33	chr07~+~128784711~128809534~	9
MRPS5	chr02~~~95752951~95787754~	8
C15orf53	chr15~+~38988798~38992239~	7
CAMK1G	chr01~+~209757044~209787284~	7
TWIST2	chr02~+~239756672~239832237~	6
FGF9	chr13~+~22245214~22278640~	5
CCDC129	chr07~+~31553684~31698334~	5
ISLR2	chr15~+~74421714~74429143~	4
C17orf51	chr17~~~21431570~21454941~	2
ADAMTS9-AS2	chr03~+~64670545~64997143~	2
GABRA5	chr15~+~27111865~27194357~	1

[000179] Also see FIG. 22, which provides detailed information regarding the 41 hydroxymethylation biomarkers of Table 4.

[000180] One preferred method for detecting the hydroxymethylation profile of a nucleic acid is described in International Patent Publication WO 2017/176630 to Quake et al., incorporated herein by reference in its entirety. That method pertains to the detection of 5-hydroxymethylcytosine patterns in cell-free DNA within the context of a sequencing scheme. An affinity tag is appended to 5hmC residues in a sample of cell-free DNA, and the tagged DNA molecules are then enriched and sequenced, with 5hmC locations identified. An illustrative example of the method, as described in Quake et al., involves initially modifying end-blunted, adaptor-ligated double-stranded DNA fragments in the cell-free sample to covalently attach biotin, as the affinity tag, to 5hmC residues. This may be carried out by selectively glucosylating 5hmC residues with uridine diphospho (UDP) glucose functionalized at the 6-position with an azide moiety, a step that is followed by a spontaneous 1,3-cycloaddition reaction with alkyne-functionalized biotin via a "click chemistry" reaction, as described previously, in Section 5, with respect to 5hmC-containing capture sequences in adapters. The DNA fragments containing the biotinylated 5hmC residues are adapter-ligated dsDNA template molecules that can then be pulled down with streptavidin beads in an "enrichment" step.

[000181] Both targeted and non-sequencing detection approaches after enrichment may also be used to quantitate specific hydroxymethylation biomarkers and loci of interest, if genome-wide coverage through shotgun sequencing is not required or desirable (generally for cost reasons). For example, after 5hmC enrichment, targeted PCR amplicons covering only

specific regions may be generated from the 5hmC-enriched templates and employed as a more narrow genome coverage approach, and used as input to sequencing or detected directly.

[000182] When a smaller number of discrete loci are of interest, the combination of these post-enrichment approaches with target amplification may also be an efficient way to reduce the number of sequencing reads (and sequencing costs) required for each sample, enabling further sample multiplexing per sequencing run and further reducing the sequencing costs required for each sample). In non-sequencing approaches, quantitative PCR or even hybridization assays could themselves be used as the quantitative readouts of the hydroxymethylation biomarkers (e.g., using direct fluorescence nucleotide labeling and microarray or other substrate capture and binding); such approaches are well known in the art, and frequently scaled to hundreds or even thousands of short amplicons.

[000183] In the present process, a 5hmC UFI sequence is added to the termini of the pulled down adapter-ligated dsDNA template molecules, so that the after amplification, pooling, and sequencing, information regarding hydroxymethylation profile can be deduced from the sequence reads obtained. That is, the sequence reads are analyzed to provide a quantitative determination of which sequences are hydroxymethylated in the cfDNA. This may be done by, e.g., counting sequence reads or, alternatively, counting the number of original starting molecules, prior to amplification, based on their fragmentation breakpoint and/or whether they contain the same molecular UFI. The use of molecular UFI sequences (or "molecular barcodes" as they are sometimes called) in conjunction with other features of the fragments (e.g., the end sequences of the fragments, which define the breakpoints) to distinguish between the fragments is known. See Casbon (2011) *Nucl. Acids Res.* 22 e81 and Fu et al. (2011) *Proc. Natl. Acad. Sci. USA* 108: 9026-31), among others. Molecular barcodes are also described in U.S. Patent Publication Nos. 2015/0044687, 2015/0024950, and 2014/0227705, and in U.S. Patent Nos. 8,835,358 and US 7,537,897, as well as a variety of other publications.

[000184] Other methods of ascertaining the hydroxymethylation profile of DNA in the cell-free nucleic sample are described in International Patent Publication WO 2019/160994 A1 to Arensdorf et al. for "Methods for the Epigenetic Analysis of DNA, particularly Cell-Free DNA" and in U.S. Patent Publication No. 2017/0298422 to Song et al., both of which are incorporated by reference herein. These references are also useful in conjunction with an embodiment of the invention in which the present combined workflow process further

includes the detection of a cfDNA methylation profile in addition to the cfDNA hydroxymethylation profile.

[000185] The Arensdorf et al. methodology described in WO 2019/160994, in the context of the present combined workflow process, can be implemented as follows:

[000186] *Dual-Biotin Technique:* After a cell-free nucleic acid sample has been extracted from a biological sample, with cfDNA having been adapter-ligated, 5hmC residues in the cfDNA are selectively labeled with an affinity tag, e.g., a biotin moiety as explained earlier herein. Biotinylation can be carried out by selective functionalization of 5hmC residues via β GT-catalyzed glucosylation with uridine diphosphoglucose-6-azide followed by a click chemistry reaction to covalently attach an alkyne-functionalized biotin moiety as explained previously. An avidin or streptavidin surface (e.g., in the form of streptavidin beads) is then used to pull out all of the dsDNA template molecules biotinylated at the 5hmC locations, which are then placed in a separate container for UFI sequence attachment during amplification. The remaining dsDNA template molecules in the supernatant are fragments that either have 5mC residues or have no modifications (the latter group including cDNA generated from cfRNA). A TET protein is then used to oxidize 5mC residues in the supernatant to 5hmC; in this case, a TET mutant protein is employed to ensure that oxidation of 5mC does not proceed beyond hydroxylation. Suitable TET mutant proteins for this purpose are described in Liu et al. (2017) *Nature Chem. Bio.* 13: 181-191, incorporated by reference herein. The β GT-catalyzed glucosylation followed by biotin functionalization is then repeated. The fragments so marked - biotinylated at each of the original 5mC locations - are pulled down with streptavidin beads. The bead-bound DNA fragments are then barcoded - with a UFI sequence than used in the first step, i.e., a 5mC UFI sequence - during amplification. Unmodified DNA fragments, i.e., fragments containing no modified cytosine residues, now remain in the supernatant. If desired, sequence-specific probes can be used to hybridize to unmethylated DNA strands. The hybridized complexes that result can be pulled out and tagged with a further UFI sequence during amplification, as before.

[000187] *Pic-Borane Methodology:* This is an alternative to the dual biotin technique, and also begins with biotinylation of 5hmC residues in adapter-ligated DNA fragments, followed by avidin or streptavidin pull-down. In this technique, however, the DNA containing unmodified 5mC residues remaining in the supernatant is oxidized beyond 5hmC, to 5caC and/or 5fC residues. Oxidation may be carried out enzymatically, using a catalytically active TET family enzyme. A "TET family enzyme" or a "TET enzyme" as those terms are used

herein refer to a catalytically active "TET family protein" or a "TET catalytically active fragment" as defined in U.S. Patent No. 9,115,386, the disclosure of which is incorporated by reference herein. A preferred TET enzyme in this context is TET2; see Ito et al. (2011) *Science* 333(6047):1300-1303. Oxidation may also be carried out chemically, using a chemical oxidizing agent. Examples of suitable oxidizing agent include, without limitation: a perruthenate anion in the form of an inorganic or organic perruthenate salt, including metal perruthenates such as potassium perruthenate (KRuO₄), tetraalkylammonium perruthenates such as tetrapropylammonium perruthenate (TPAP) and tetrabutylammonium perruthenate (TBAP), and polymer supported perruthenate (PSP); and inorganic peroxo compounds and compositions such as peroxotungstate or a copper (II) perchlorate / TEMPO combination. It is unnecessary at this point to separate 5fC-containing fragments from 5caC-containing fragments, insofar as in the next step of the process, both 5fC residues and 5caC residues are converted to dihydrouracil (DHU).

[000188] That is, following oxidation of 5mC residues to 5fC and 5caC, an organic borane is added to reduce, deaminate, and either decarboxylate or deformylates the oxidized 5mC residues. The resulting dsDNA template molecules contain DHU in place of the original 5mC residues, and can be amplified, pooled, and sequenced, along with other dsDNA template molecules deriving from the same sample.

[000189] The organic borane may be characterized as a complex of borane and a nitrogen-containing compound selected from nitrogen heterocycles and tertiary amines. The nitrogen heterocycle may be monocyclic, bicyclic, or polycyclic, but is typically monocyclic, in the form of a 5- or 6-membered ring that contains a nitrogen heteroatom and optionally one or more additional heteroatoms selected from N, O, and S. The nitrogen heterocycle may be aromatic or alicyclic. Preferred nitrogen heterocycles herein include 2-pyrroline, 2*H*-pyrrole, 1*H*-pyrrole, pyrrolidine, imidazolidine, 2-pyrazoline, 2-imidazoline, pyrazole, imidazole, 1,2,4-triazole, 1,2,4-triazole, pyridazine, pyrimidine, pyrazine, 1,2,4-triazine, and 1,3,5-triazine, any of which may be unsubstituted or substituted with one or more non-hydrogen substituents. Typical non-hydrogen substituents are alkyl groups, particularly lower alkyl groups, such as methyl, ethyl, n-propyl, isopropyl, n-butyl, isobutyl, t-butyl, and the like. Exemplary compounds include pyridine borane, 2-methylpyridine borane (also referred to as 2-picoline borane), and 5-ethyl-2-pyridine. Further information concerning these organic boranes and reaction thereof to convert oxidized 5mC residues to DHU may be found in the Arensdorf patent publication cited above.

[000190] *Biotin/Native 5mC Enrichment Method:* This is an alternative to the dual biotin technique, and begins with biotinylation of 5hmC residues in adapter-ligated DNA fragments, followed by avidin or streptavidin pull-down. Here, however, instead of modifying the methylated DNA that remains in the supernatant, an anti-5mC antibody or an MBD protein is used to capture and pull down native 5mC-containing fragments. This technique is less preferred herein, insofar as it does not result in the generation of dsDNA template molecules that can be amplified, pooled, and sequenced with other dsDNA template molecules deriving from the same sample.

[000191] **3. Methods of use:**

[000192] As explained in the preceding section, the invention, in one embodiment, provides a method for predicting the risk that a patient with an identified pancreatic lesion has pancreatic cancer. Also provided are diagnostic, prognostic, and predictive uses of hydroxymethylation profiles, as well as uses in patient monitoring, evaluation of treatment options, and evaluation of treatment efficacy, wherein, in each method of use, the hydroxymethylation profile generated is combined with clinical parameters and optionally with one or more other risk factors in each method of use. All of the methods involve the generation of a hydroxymethylation profile comprising measurements of hydroxymethylation levels at each of a plurality of hydroxymethylation biomarker loci.

[000193] Among the provided diagnostic, prognostic, and predictive methods are those which employ statistical analysis and biomathematical algorithms and predictive models to analyze the detected hydroxymethylation information. Some embodiments include methods and systems for analyzing the hydroxymethylation information in classification, staging, prognosis, treatment design, evaluation of treatment options, prediction of outcomes (e.g., predicting development of metastases), and the like.

[000194] Also provided are methods that use evaluation of hydroxymethylation levels at the biomarker loci in treatment development and patient monitoring, including evaluation of a patient's response to treatment and patient-specific or individualized treatment strategies. In some embodiments, the methods are used in conjunction with treatment, for example, by generating a hydroxymethylation profile weekly or monthly before and/or after treatment. As the hydroxymethylation levels at certain biomarker loci correlate with the progression of disease, ineffectiveness or effectiveness of treatment, and/or the recurrence or lack thereof of disease, the regular generation of hydroxymethylation profiles within an extended monitoring or treatment period is useful. In some aspects, the information obtained may indicate that a

different treatment strategy is preferable. Thus, provided herein are therapeutic methods, in which biomarker evaluation is performed prior to treatment, and then used to monitor therapeutic effects.

[000195] More specifically, at various points in time after initiating or resuming treatment, significant changes in hydroxymethylation levels at one or more of the biomarker loci may be seen, indicating that a therapeutic strategy is or is not successful, that disease is recurring, or that a different therapeutic approach should be used. In some embodiments, the therapeutic strategy is changed following a hydroxymethylation analysis, such as by adding a different therapeutic intervention, either in addition to or in place of a prior approach, by increasing or decreasing the aggressiveness or frequency of the approach, or by stopping or reinstituting a treatment regimen.

[000196] In another aspect, the hydroxymethylation levels at each of the biomarker loci are used to identify the presence of pancreatic cancer or a risk of developing pancreatic cancer for the first time.

[000197] In some aspects, the methods determine whether or not the assayed patient is responsive to treatment, such as a subject who is clinically categorized as in complete remission or exhibiting stable disease. In some aspects, methods are provided for distinguishing treatment-responsive and non-responsive patients, and for distinguishing patients with stable disease or those in complete remission, and those with progressive disease.

[000198] In various aspects, the methods and systems make such calls with at least at or about 65, 70, 75, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, or 100% correct call rate (i.e., accuracy), specificity, or sensitivity.

[000199] All of the aforementioned methods are encompassed by the present invention. Preferred methods herein include, without limitation:

[000200] a method for evaluating the risk that an identified pancreatic lesion in a patient is cancerous;

[000201] a method for monitoring an identified pancreatic lesion in a patient, which involves an analysis of hydroxymethylation changes over time;

[000202] a method for managing a patient with an identified pancreatic lesion, which involves an evaluation of treatment options based on a hydroxymethylation profile;

[000203] a method for monitoring the effectiveness of treatment in a patient with an identified pancreatic lesion, which involves an analysis of hydroxymethylation profiles generated at selected time intervals within an extended monitoring period;

[000204] a method for reducing unnecessary surgical resection of a pancreatic lesion by evaluating the risk that the pancreatic lesion is cancerous using a hydroxymethylation profile; and

[000205] a method for identifying the risk that a patient without an identified pancreatic lesion will develop pancreatic cancer.

[000206] **4. Statistical analyses, mathematical algorithms and predictive models:**

[000207] Typically, the methods of the invention include statistical analysis and mathematical modeling used to analyze high-dimensional and multimodal biomedical data, such as the data obtained using the present methods for generating and comparing hydroxymethylation profiles. More specifically, the methods make use of one or more objective algorithms, models, and analytical methods that include mathematical analyses based on topographic, pattern-recognition based protocols, e.g., support vector machines (SVM), linear discriminant analysis (LDA), naive Bayes (NB), and K-nearest neighbor (KNN) protocols, as well as other supervised learning algorithms and models, such as Decision Tree, Perceptron, and regularized discriminant analysis (RDA), and similar models and algorithms well-known in the art (Gallant S I, "Perceptron-based learning algorithms," Perceptron-based learning algorithms 1990; 1(2):179-91).

[000208] Statistical analyses include determining mean (M), e.g., geometric mean, standard deviations (SD), Geometric Fold Change (FC), and the like. Whether differences in hydroxymethylation levels are deemed significant may be determined by well-known statistical approaches, typically by designating a threshold for a particular statistical parameter, such as a threshold p-value (e.g., $p < 0.05$), a threshold S-value (e.g., ± 0.4 , with $S \leq -0.4$ or $S > 0.4$), or other value, at which differences are deemed significant, for example when the level of biomarker hydroxymethylation in a hydroxymethylation profile is considered significantly increased or decreased, respectively, relative to the hydroxymethylation level at the same hydroxymethylation biomarker locus in a reference hydroxymethylation profile.

[000209] In one aspect, the methods of the invention apply the mathematical formulations, algorithms or models to distinguish between normal and cancerous samples, and between various sub-types, stages, and other aspects of disease or disease outcome. In another aspect,

the methods are used for prediction, classification, prognosis, and treatment monitoring and design.

[000210] For the comparison of hydroxymethylation levels or other values, data are compressed. Compression typically is by Principal Component Analysis (PCA) or a similar technique for visualizing the structure of high-dimensional data. PCA is used to reduce dimensionality of the data (e.g., measured expression values) into uncorrelated principal components (PCs) that explain or represent a majority of the variance in the data, such as about 50, 60, 70, 75, 80, 85, 90, 95 or 99% of the variance. PCA allows the visualization of biomarker levels and the comparison of hydroxymethylation profiles, such as between normal or reference samples and test samples. PCA mapping, e.g., 3-component PCA mapping is used to map data to a three-dimensional space for visualization, such as by assigning first, second, and third PCs to the x-, y-, and z-axes, respectively.

[000211] In some embodiments, there is a linear correlation between hydroxymethylation levels of two or more biomarkers. Pearson's Correlation (PC) coefficients may be used to assess linear relationships (correlations) between pairs of values, such as between hydroxymethylation levels of a biomarker. This analysis may be used to linearly separate distribution in expression patterns, by calculating PC coefficients for individual pairs of the biomarkers (plotted on x- and y-axes of individual Similarity Matrices). Thresholds may be set for varying degrees of linear correlation, such as a threshold for highly linear correlation of ($R_{sup.2} > 0.50$, or 0.40). Linear classifiers can be applied to the datasets. In one example, the correlation coefficient is 1.0.

[000212] In some embodiments, Feature Selection (FS) is applied to remove the most redundant features from a dataset, such as a hydroxymethylation biomarker dataset. FS enhances the generalization capability, accelerates the learning process, and improves model interpretability. In one aspect, FS is employed using a "greedy forward" selection approach, selecting the most relevant subset of features for the robust learning models. (Peng H, Long F, Ding C, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005; 27(8):1226-38). In some embodiments, SVM algorithms are used for classification of data by increasing the margin between the n data sets (Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge: Cambridge University Press, 2000).

[000213] Analytic classification of the hydroxymethylation biomarkers herein can be made according to predictive modeling methods that set a threshold for determining the probability that a sample (e.g., a cfDNA sample obtained from a patient) belongs to a given class (e.g., elevated risk of developing pancreatic cancer). The probability preferably is at least 50%, or at least 60%, or at least 70%, or at least 80% or higher. Classifications also can be made by determining whether a comparison between an obtained dataset and a reference dataset yields a statistically significant difference. If so, then the sample from which the dataset was obtained is classified as not belonging to the reference dataset class. Conversely, if such a comparison is not statistically significantly different from the reference dataset, then the sample from which the dataset was obtained is classified as belonging to the reference dataset class.

[000214] The predictive ability of a model can be evaluated according to its ability to provide a quality metric, e.g. AUROC (area under the ROC curve) or accuracy, of a particular value, or range of values. Area under the curve measures are useful for comparing the accuracy of a classifier across the complete data range. Classifiers with a greater AUC have a greater capacity to classify unknowns correctly between two groups of interest. In some embodiments, a desired quality threshold is a predictive model that will classify a sample with an accuracy of at least about 0.7, at least about 0.75, at least about 0.8, at least about 0.85, at least about 0.9, at least about 0.95, or higher. As an alternative measure, a desired quality threshold can refer to a predictive model that will classify a sample with an AUC of at least about 0.7, at least about 0.75, at least about 0.8, at least about 0.85, at least about 0.9, or higher.

[000215] As is known in the art, the relative sensitivity and specificity of a predictive model can be adjusted to favor either the selectivity metric or the sensitivity metric, where the two metrics have an inverse relationship. The limits in a model as described above can be adjusted to provide a selected sensitivity or specificity level, depending on the particular requirements of the test being performed. One or both of sensitivity and specificity can be at least about 0.7, at least about 0.75, at least about 0.8, at least about 0.85, at least about 0.9, at least about 0.95, at least about 0.98, at least about 0.99, or higher.

[000216] Raw data can be initially analyzed by measuring the hydroxymethylation level for each biomarker. The data can be manipulated, for example, raw data can be transformed using standard curves, and the average of multiple measurements, if made, can be used to calculate the average and standard deviation for each patient. The data are then input into a

selected predictive model, which will classify the sample. The resulting information can be communicated to a patient or health care provider, usually in the form of a written report.

[000217] To generate a predictive model for pancreatic cancer, a robust data set, comprising known control samples and samples corresponding to pancreatic cancer, is used in a training set. A sample size can be selected using generally accepted criteria. As discussed above, different statistical methods can be used to obtain a highly accurate predictive model. The examples herein provide representative such analyses.

[000218] In one embodiment, hierarchical clustering is performed in the derivation of a predictive model, where the Pearson correlation is employed as the clustering metric. One approach is to consider a dataset as a "learning sample" in a problem of "supervised learning." CART is a standard in applications to medicine (Singer, *Recursive Partitioning in the Health Sciences* (Springer, 1999)) and can be modified by transforming any qualitative features to quantitative features, sorting them by attained significance levels, and a selected regularization method then applied (e.g., Elastic Net or Lasso).

[000219] In some embodiments, the predictive models include Decision Tree, which maps observations about an item to a conclusion about its target value (Zhang et al., "*Recursive Partitioning in the Health Sciences*," in *Statistics for Biology and Health* (Springer, 1999.)). The leaves of the tree represent classifications and branches represent conjunctions of features that devolve into the individual classifications.

[000220] The predictive models and algorithms may further include Perceptron, a linear classifier that forms a feed forward neural network and maps an input variable to a binary classifier (Gallant (1990), "Perceptron-based learning algorithms," in *IEEE Transactions on Neural Networks* 1(2):179-191). In this model, the learning rate is a constant that regulates the speed of learning. A lower learning rate improves the classification model, while increasing the time to process the variable (Markey et al. (2002) *Comput Biol Med* 32(2):99-109).

[000221] These and other aspects of the invention are further described and illustrated by way of the following examples.

EXAMPLE 1

[000222] (a) *Study design and clinical cohort:*

[000223] Plasma specimens from subjects without or with pancreatic ductal adenocarcinoma were collected at multiple institutions in different geographic regions of the United States and Germany. This study group, Study Group 1, included 41 PDAC patients

and 51 non-cancer subjects. These PDAC and non-cancer patient samples satisfied the study inclusion criteria, which included a minimum subject age of 18 years as well as confirmed pathologic diagnosis of adenocarcinoma of any subtype at the time of surgical resection, for subjects in the cancer cohort. The non-cancer cohort was identified as satisfying the study inclusion criteria and patients were specifically negative for any form of cancer. Neither cohort was being treated with medication for disease at the time of blood collection. There were no statistically significant differences in subject age or gender between the two cohorts, but there was a statistically significant greater tobacco exposure in the PDAC cohort, as expected given smoking is a common risk factor for pancreatic cancer. The clinical characteristics of the cancer and non-cancer cohorts are set forth in Table 5:

[000224] Table 5:

	No Cancer	Cancer
Age⁺		
	66.0	71.2
Gender(%)		
Male	60.0	45.1
Smoking History		
Status(%)		
Current	19.5	19.6
Former	29.3	37.3
None	51.2	43.1
Pack-Years⁺		
Current	5.3	29.6
Former	20.5	24.2
None	NA	NA
Pack-Years⁺		
All	14.4	25.7
Time Since Cessation⁺		
Months	264.2	272.3
Stage		
I	NA	18
II	NA	61
III	NA	7.8
IV	NA	14

+mean of Non-Cancer and Cancer

Other values are percentages of each category in the Non-Cancer and Cancer

groups.

[000225] (b) Sequencing results and metrics:

[000226] 5hmC-enriched libraries were prepared using the cell-free "5hmC-Seal" method described in International Patent Publication WO 2017/176630 to Quake et al., Song et al.

(2011) 29: 68-72, and Han et al. (2016) Mol. Cell 63:711-19, the disclosures of which are incorporated by reference herein. Briefly, hMe-Seal is a low-input, whole-genome cell-free 5hmC sequencing method based on selective chemical labeling, in which β -glucosyltransferase is used to selectively label 5hmC with a biotin moiety via an azide-modified glucose for pull-down of 5hmC-containing DNA fragments for sequencing. In implementing hMe-Seal in the present case, the cfDNA was first ligated with sequencing adapters, followed by selective labeling of 5hmC with β -GT, and affinity enrichment via selective pull-down of DNA fragments containing biotin-labeled 5hmC using streptavidin beads. PCR was then carried out directly from the beads (i.e., instead of eluting the captured DNA) to minimize sample loss during purification. A median number of unique read pairs of 9.1 and 10.7 million in the PDAC and non-cancer cohorts, respectively, were produced. Filtering criteria to enable the determination of high quality 5hmC libraries were established from previous studies (Fonseca et al. (2018), supra), yielding 51 in the pancreatic cancer group and 41 in the non-cancer group. Extensive analysis did not reveal batch processing effects occurring specifically in either study cohort.

[000227] (c) Cohort-based distributions of 5hmC densities into functional regions:

[000228] The vast majority of 5hmC loci, as measured by increased read density and detected by MACS2 as peaks, were found to occur, on average, in non-coding, intragenic regions of the genome, i.e., intronic, transposon repeats - SINEs and LINEs, and intergenic, as illustrated in FIG. 3, with no preferential 5hmC distribution in any one disease cohort. These regions displayed low 5hmC enrichment (intron, FIG. 4) or even depletion of 5hmC sites (intergenic and LINE elements, FIG. 4). Instead, 5hmC enrichment occurred more frequently in promoters, UTRs, exons, transcription termination sites (TTS) and SINE elements as measured relative to the genome background. Significant differences in the enrichment of 5hmC peaks in functional regions were observed in a disease cohort specific manner. Increases in enrichment in PDAC were measured in exons, 3'UTR, and TTS, whereas decreases were found in promoter and LINEs, which themselves were either 5hmC-enriched or 5hmC-depleted, respectively (FIG. 5). These global changes were found to occur in a statistically significant manner in each cohort and were also found to occur in a cancer stage specific manner, with gradual increases (exon, 3'UTR and TTS) or decreases (promoter and LINE) in later stage patients (FIG. 6).

[000229] Chemical modifications like methylation and acetylation on histone proteins were inferred in relation to 5hmC occupancy, using the existing histone maps from PANC-1 cell lines (see LeRoy et al. (2013) *Epigenetics & Chromatin* 6:20). Notably, 5hmC decreases were seen in PDAC coincident with H3K4Me3 loci, a transcriptional activating mark (but not H3K4Me1), combined with 5hmC decreases in H3K27Ac and H3K27Me3 loci, which are transcription activating and inactivating respectively (FIG. 7). The statistically significant changes in H3K4Me3, H3K27Ac and H3K27Me3 all exhibited an ongoing reduction in later stage PDAC patients compared with the non-cancer cohort. The H3K27Ac mark had the largest density of 5hmC occupancy in both the cancer and non-cancer cohort and the highest similarity to the Panc1 cell line histone map (FIG. 8A). Conversely, H3K27Me3 exhibited the lowest density of 5hmC occupancy in both cohorts and the lowest similarity to the PANC-1 cell line histone map (FIG. 8B)

[000230] *(d) Identification of disease specific genes from plasma samples:*

[000231] Differential analysis of 5hmC densities in genes revealed 6,496 and 6,684 genes with an increased and decreased 5hmC density, respectively, in PDAC, compared to non-cancer samples (FIG. 11). Further filtering of this gene set (fold change $\geq |1.5|$ in PDAC versus non-cancer and average log 2 CPM ≥ 4 counts, 142 genes total) revealed annotated genes with increased 5hmC density and whose biology is related to pancreas development (GATA4, GATA6, PROX1, ONECUT1) and/or implicated in cancer (YAP1, TEAD1, PROX1, ONECUT2/ONECUT1, IGF1 and IGF2). Inspection of the Molecular Signatures Database (MSigDB) for relevant pathways comprising the 142 genes with enriched 5hmC densities revealed a numeric preponderance of pathways down-regulated in liver cancer (5 of the top 10 most significant pathways, as indicated in Table 6). The differential representation analysis coupled with filtering (fold change $\geq |1.5|$ in PDAC versus non-cancer and log CPM of 5hmC ≥ 4) also revealed 178 genes with a decreased 5hmC density in pancreatic cancer cfDNA (Table 7). Closer inspection of these pathways with decreased 5hmC representation revealed fundamental pathways in immune system regulation (3 of the top 10 most significant pathways, as may be seen in Table 6). Pancreatic cancers are typically diagnosed at late stage where disease prognosis is poor, as exemplified by a 5-year survival rate of 8.2%. Earlier diagnosis would be beneficial by enabling surgical resection or earlier application of therapeutic regimens. The above example illustrates that pancreatic adenocarcinoma can be detected in a non-invasive manner by interrogating changes in 5-hydroxymethylation cytosine status of circulating cell free DNA in the plasma of a PDAC

cohort in comparison with a non-cancer cohort. The inventors found that 5hmC sites are enriched in a disease-specific and stage-specific manner in exons, 3'UTRs, and transcription termination sites.

[000232] Expanding gene set enrichment analysis to include the full data set of all genes revealed that more than 30% of immune related pathways have reduced 5-hydroxymethylation across early and late stage PDAC (FIG. 10). Principal component analysis (PCA) using either the 13,180 genes with statistically significant variation in 5hmC counts (FIG. 11), or the 320 genes filtered at the extremes of 5hmC representation in PDAC (FIG. 12), revealed partitioning of the PDAC samples from the non-cancer samples equally well, indicating no loss of partitioning signal using a biologically relevant and statistically filtered gene set.

Table 6: Top ten pathways represented by 142 genes with increased 5hmC density in PDAC samples versus non-cancer samples (also see Collin et al. (2018), "Detection of Early Stage Pancreatic Cancer Using 5-Hydroxymethylcytosine Signatures in Circulating Cell-Free DNA," *bioRxiv*, doi:<https://dx.doi.org/10.1101/422675>, incorporated by reference herein):

Gene Set*	Description	k/K	p-value	FDR q-value
1	Genes down-regulated in liver tissue upon knockout of HNF1A (GeneID=6927).	0.0892	3.92E-17	5.61E-13
2	Genes down-regulated in tumor compared to non-tumor liver samples from patients with hepatocellular carcinoma (HCC).	0.2041	2.35E-16	1.68E-12
3	The chemical reactions and pathways involving small molecules (any low molecular weight, monomeric, non-encoded molecule).	0.0175	5.53E-16	2.64E-12
4	Liver-selective genes.	0.0615	7.06E-16	2.72E-12
5	Genes from subtype S3 signature of HCC: hepatocyte differentiation.	0.0564	2.73E-15	7.81E-12
6	Genes down-regulated in liver tumor compared to the normal adjacent tissue.	0.0547	4.22E-15	1.01E-11
7	The chemical reactions and pathways involving lipids and lipidic molecules (fatty alcohols, sphingoids, phospholipids, glycolipids, sterols, etc.).	0.0216	5.47E-15	1.12E-11
8	The chemical reactions and pathways involving organic acids.	0.0241	7.39E-15	1.32E-11
9	Any process that results in a change in state or activity of a cell or an organism (movement, secretion, enzyme production, gene expression, etc.) as a result of a stimulus arising within the organism.	0.0179	1.08E-13	1.61E-10
10	Down-regulated genes distinguishing between early gastric cancer (EGC) and normal tissue samples.	0.0409	2.98E-13	4.27E-10

*Gene set names:

1: SERVITJA_LIVER_HNF1A_TARGETS_DN

2: LEE_LIVER_CANCER

3: GO_SMALL_MOLECULE_METABOLIC_PROCESS

- 4: HSIAO_LIVER_SPECIFIC_GENES
- 5: HOSHIDA_LIVER_CANCER_SUBCLASS_S3
- 6: ACEVEDO_LIVER_TUMOR_VS_NORMAL_ADJACENT
- 7: GO_LIPID_METABOLIC_PROCESS
- 8: GO_ORGANIC_ACID_METABOLIC_PROCESS
- 9: GO_RESPONSE_TO_ENDOGENOUS_STIMULUS
- 10: VECCHI_GASTRIC_CANCER_EARLY_DN

Table 7: Top ten pathways represented by 178 genes with decreased 5hmC density in PDAC samples versus non-cancer samples (also see Collin et al. (2018), *supra*):

Gene Set*	Description	k/K	p-value	FDR q-value
1	Genes involved in hemostasis	0.0708	1.80E-33	2.58E-29
2	Any process that modulates the frequency, rate, or extent of an immune system process.	0.0314	1.06E-29	7.56E-26
3	Genes down-regulated in CD34+ (GeneID=947) cells by intermediate activity levels of STAT5A (GeneID=6776); predominant long-term growth and self-renewal phenotype.	0.108	9.52E-28	4.54E-24
4	Any process involved in the development or functioning of the immune system.	0.0242	1.62E-27	5.79E-24
5	Any process that modulates the levels of body fluids.	0.0553	1.76E-25	5.05E-22
6	A change in the morphology or behavior of a cell resulting from exposure to an activating factor such as a cellular or soluble ligand.	0.0511	2.19E-25	5.23E-22
7	Any process that activates or increases the frequency, rate, or extent of an immune system process.	0.0381	8.38E-25	1.71E-21
8	Any process that modulates the frequency, rate, or extent of cell activation.	0.0558	1.11E-24	1.98E-21
9	Genes involved in platelet activation, signaling, and aggregation.	0.0962	3.27E-23	5.19E-20
10	Any process that modulates the frequency, rate, or extent of attachment of a cell to another cell or to the extracellular matrix.	0.0429	1.04E-21	1.49E-18

*Gene set names:

- 1: REACTOME_HEMOSTASIS
- 2: GO_REGULATION_OF_IMMUNE_SYSTEM_PROCESS
- 3: WIERENGA_STATSA_TARGETS_DN
- 4: GO_IMMUNE_SYSTEM_PROCESS
- 5: GO_REGULATION_OF_BODY_FLUID_LEVELS
- 6: GO_CELL_ACTIVATION
- 7: GO_POSITIVE_REGULATION_OF_IMMUNE_SYSTEM_PROCESS
- 8: GO_REGULATION_OF_CELL_ACTIVATION
- 9: REACTOME_PLATELET_ACTIVATION_SIGNALING_AND_AGGREGATION
- 10: GO_REGULATION_OF_CELL_ADHESION

[000233] Regularized regression models were built using 5hmC densities in statistically filtered genes or a comprehensive set of highly variable gene counts and performed with an

AUC = 0.94-0.96 on training data. The inventors tested the capability of classifying PDAC and non-cancer samples with the elastic net and lasso models on two external pancreatic cancer 5hmC data sets and found validation performance to be AUC = 0.74-0.97. The findings show that 5hmC changes enable classification of PDAC patients with high fidelity.

[000234] (e) Predictive models for the detection of pancreatic cancer in cfDNA:

[000235] Regularized logistic regression analysis was performed in order to determine whether gene-based features are present in the PDAC and non-cancer cohorts that enable the classification of patient samples. The full set of 92 patient samples was partitioned into a training and test set comprising 75% and 25% of the patient data, respectively, and 65% of the genes with the most variable 5hmC count was employed for model selection. Two methods of regularization were used, Elastic Net (glmnet) and Lasso (glmnet2) (Yu et al. (2016) *BMC Bioinformatics* 17:108).

[000236] Both regularization methods require the specification of hyper-parameters that control the level of regularization used in the fit. Hyper-parameters were selected based on out-of-fold performance on 30 repetitions of 10-fold cross-validated analysis of the training data. Out-of-fold assessments were based on the samples in the left-out fold at each step of the cross-validated analysis. The training set yielded an out-of-fold performance metric, Area Under Curve (AUC), of 0.96 (elastic net and lasso) with an internal sample test AUC of 0.84 (elastic net) and 0.88 (lasso) (FIG. 15). The distribution of probability scores shows that within training data, both models classify well (FIG. 16) but that marginally fewer misclassified samples are found with the elastic net model when the specificity is set at 75%, i.e., fewer cancer samples score below the third quartile non-cancer score and fewer non-cancer samples score above the same third quartile non-cancer score. Next, the training model was tested on an external validation set of patient samples. These include pancreatic cancer samples from Li et al. (2017) *Cell Research* 27:1243-1257 (pancreas subtype not specified; 23 subjects with pancreatic cancer, 53 healthy) and Song et al. (2017) *Cell Research* 27: 1231-42 (pancreas subtype specified as adenocarcinoma; 7 subjects with pancreatic cancer, 10 healthy). The validation set exhibited a performance with AUC = 0.78 (elastic net and lasso) in the Li et al data and AUC = 0.99 (elastic net) and 0.97 (lasso) in the Song et al. (2017) data (FIG. 12).

[000237] The effect of feature selection on prediction performance was evaluated by filtering the initial set of significant genes (FIG. 10) to satisfy a 1.5-fold differential 5hmC representation in the PDAC cohort with median representation of gene counts of log2 average

5hmC representation > 4. The same regularized regression models were built using this set of 287 genes with increased 5hmC and 343 genes with decreased 5hmC counts, employing a similar setup for training and testing as defined previously (75% data used for training, 25% data used for test) and found training set AUC = 0.96 (elastic net) and 0.94 (lasso). Not surprisingly, internal testing yields a high performance with AUC = 0.92 (elastic net) and 0.93 (lasso). Of greater interest was the performance on external data sets with AUC = 0.74 (elastic net) and 0.67 (lasso) for Li et al data and AUC = 0.97 (elastic net) and 0.94 (lasso) for Song et al. (2017) data. This suggests that genes with evident enrichment of biological signal relevant to pancreatic cancer and/or pancreas development do not perform much better than an algorithmically driven selection of features during regression training, as has been shown elsewhere (*ibid.*). Hierarchical clustering of these significant genes (287 with increased 5hmC +343 with decreased 5hmC) showed good partitioning of the pancreatic cancer samples in the Stanford data set but less pronounced separation of Chicago data (FIG. 16).

[000238] The final models fitted to the 65% most variable 5hmC gene features, using hyper-parameter values determined from the training set data analysis, were fitted to the whole cohort of PDAC and non-cancer samples and this yielded models with 109 genes (elastic net) and 47 genes (lasso). The models were found to possess t-scores that are concordant both the Li et al and Song et al. (2017) data sets (FIG. 17).

[000239] *Discussion:*

[000240] The experimental work detailed above focused on the discovery of cfDNA-specific hydroxymethylation-based biomarkers that would facilitate the development of molecular diagnostic tests to detect pancreatic cancer at earlier stages. The data discussed above and presented in the figures highlight the ability to detect differentially hydroxymethylated genes whose underlying biology shows association with both pancreas and cancer development as well as established trends in marked, known functional regions of the genome. Furthermore, by using either 5hmC signals from biologically significant genes or from regularized regression methods, predictive models can be built with AUC = 0.94-0.96 with an external data set validation AUC = 0.74-0.97 (elastic net models).

[000241] The 5hmC signal was found to be enriched in gene-centric sequence types (promoter, exons, UTR and TTS), as well as transposable elements like SINEs (enriched) and LINEs (depleted) (FIGS. 3 and 4). These hydroxymethylation changes in functional regions have been reported in cfDNA from colorectal, esophageal, liver, and lung cancer (see Li et al. (2017), *supra*; Tian et al. (2018) *Cell Res* 5:597-600; Cai et al. (2018), "5-

Hydroxymethylcytosines from Circulating Cell-free DNA as Diagnostic and Prognostic Markers....," *bioRxiv* (doi:<https://doi.org/10.1101/424978>), and Zhang et al. (2018) *Genomics, Proteomics & Bioinformatics* 16: 187-199); however, no PDAC-specific gains or losses in hydroxymethylation were observed in functional regions. In addition to enrichment and depletion of 5hmC in functional regions, there was a novel PDAC specific 5hmC increase in exons, TTS and 3'UTR and a 5hmC decrease in promoters and LINE elements (FIG. 5). In ES cells, the decrease of 5-hydroxymethylation in the promoter region has been shown to associate with gene transcription (see Szulwach et al. (2011) *PLoS Genetics* 7(6): e1002154). An increase in disease relevant transcription is be implicitly supported in the above data by the 5hmC increase seen in gene-centric features and the apparent decreasing trend of 5hmC in promoter regions toward late stage PDAC (FIG. 6).

[000242] Dynamic changes in chromatin have been shown to control cell development and transition of cells with oncogenic potential; see Bernhart et al. (2016) *Scientific Reports* 6, Article number 37393. The PDAC specific decrease of 5hmC in H3K4me3 loci appears to be coincident with a non-statistically significant increase of 5hmC in H3K4me1 (FIG. 7). These DNA hydroxymethylation patterns complement each other both in genomic location and the histone marks they occupy (FIGS 8A and 8B), and also suggest disease-specific increases in gene transcription via chromatin modifications, given the known permissive transcriptional function associated with H3K4me3/me1. An intriguing aspect of the precision of 5hmC patterns in these regions of known functional sequence suggest a widespread function for hydroxylation in the epigenetic control of transcriptional processes.

[000243] In this study, genes were identified whose increased 5hmC signals in highlighted pathways are implicated in liver cancer (Table 7). MSigDB does not currently contain pathways annotated for pancreatic cancer; see Subramanian et al. (2005), *PNAS* 102:15545-50. Two approaches were used for gene set enrichment analysis, either using genes with significantly decreased 5hmC or performing GSEA on all reporting genes. The results indicated that close to one third of immune system pathways were implicated in pancreatic cancer. Assuming the strong association between 5hmC extent and gene transcription, this result suggests that immune system function is decreased in PDAC patients. Inspection of individual genes that were either significantly increased or decreased in functional regions reveals genes implicated in normal pancreas development, for instance the transcription factors GATA4, GATA6, PROX1, ONECUT2, and also genes whose increased expression is implicated in cancer like YAP1, TEAD, PROX1, ONECUT2, ONECUT1, IGF1 and IGF2.

[000244] Using genes whose 5hmC densities are significantly changed in PDAC, with annotated relevant biology, it was possible to build regularized regression models whose performance matched model building using algorithm gene based selection. This provided confidence that the models used, whose performance is high (training AUC = 0.94-0.96 with an external data set validation AUC = 0.74-0.97), is measuring underlying biological signals relevant to PDAC. Despite the large number of significantly hydroxymethylated genes, the regularized regression models selected 100 genes or fewer. However, the fact that 13,180 significantly represented genes were detected provide evidence that other biological signals may also reside in the data set. Smoking status is a known risk factor for PDAC up to 20 years post-smoking cessation, and DNA methylation changes have been associated with tobacco-based toxins (Lee (2013) *Front Genet* 4:132). In a retrospective case-control designed study, smokers constituted 59% and 49% of PDAC and non-cancer cohorts respectively, indicating that smokers are equally spread in each cohort. Consequently, the association of smoking in the PDAC cohort could not have accounted for the significantly hydroxymethylated genes found. However, a more extensive study focused on sub-partitioning PDAC and non-cancer patient into never and ever smokers with pack-year characteristics would enable us to answer the impact of smoking on the hydroxymethylome in PDAC patients.

EXAMPLE 2

[000245] Example 1 was repeated with an additional study group, Study Group 2, of 41 PDAC and 82 non-cancer subjects. The clinical characteristics of the cancer and non-cancer cohorts in Study Group 2 are set forth in Table 8.

[000246] Table 8:

	No Cancer	Cancer
Age⁺		
	66.0	65.5
Gender(%)		
Male	50	44
Smoking History		
Status(%)		
Current	5	12
Former	54	39
Never	41	49
Stage		
NA	NA	5

I	NA	22
II	NA	29
III	NA	15
IV	NA	29

+mean of Non-Cancer and Cancer

[000247] The procedures documented in Example 1 were followed to generate the 611 hydroxymethylation biomarkers set forth in FIG. 21.

EXAMPLE 3

[000248] Example 1 was repeated with a further study group, Study Group 3, of 53 PDAC and 53 non-cancer subjects. The clinical characteristics of the cancer and non-cancer cohorts in Study Group 3 are set forth in Table 9.

[000249] Table 9:

	No Cancer	Cancer
Age⁺		
	66.0	66.4
Gender(%)		
Male	44	53
Smoking History		
Status(%)		
Current	21	21
Former	32	32
Never	47	47
Stage		
NA	NA	4
I	NA	21
II	NA	36
III	NA	11
IV	NA	28

+mean of Non-Cancer and Cancer

[000250] The procedures documented in Example 1 were followed to generate the 41 hydroxymethylation biomarkers set forth in Table 4, provided earlier herein, and in FIG. 22.

CLAIMS:

1. A method for evaluating the risk that an identified pancreatic lesion in a patient is cancerous, the method comprising:

- (a) obtaining a cell-free DNA sample from the patient;
- (b) enriching for hydroxymethylated DNA in the sample;
- (c) quantifying the nucleic acids in the enriched sample that map to each of a plurality of selected loci in a reference hydroxymethylation profile, wherein each selected locus comprises a hydroxymethylation biomarker;
- (d) comparing, at each locus, the hydroxymethylation level of the sample with the hydroxymethylation level in the reference profile, to ascertain differences in hydroxymethylation levels between the sample and the reference profile for each biomarker; and
- (e) calculating an index value representing the risk that the pancreatic lesion is cancerous from the comparison in step (d) combined with at least one additional parameter correlated with the risk that an individual has pancreatic cancer.

2. The method of claim 1, wherein the plurality of additional parameters is selected from lesion size; lesion location; presence or absence of pancreatic inflammation; jaundice; presence or absence of other symptoms; patient age; weight; gender; ethnicity; family history; genetic mutations; diabetes; physical activity; diet; pro-inflammatory cytokine levels; and smoking status of the patient.

3. The method of claim 1, wherein the reference hydroxymethylation profile represents a composite of a plurality of hydroxymethylation profiles for individuals who have not had a pancreatic lesion.

4. The method of claim 3, wherein the patient has a risk factor for pancreatic cancer and the reference hydroxymethylation profile represents a composite of a plurality of hydroxymethylation profiles for individuals who have the risk factor.

5. The method of claim 4, wherein the risk factor is pancreatitis and the reference hydroxymethylation profile represents a composite of a plurality of hydroxymethylation profiles for individuals who have been diagnosed with pancreatitis.

6. The method of claim 4, wherein the pancreatic lesion was identified by imaging, and the reference hydroxymethylation profile represents a composite of a plurality of hydroxymethylation profiles for individuals who have had a pancreatic lesion identified on an imaging scan.

7. The method of claim 1, wherein the cell-free DNA sample is extracted from a blood sample.

8. The method of claim 1, wherein the cell-free DNA sample is extracted from pancreatic cyst fluid.

9. The method of claim 1, further comprising generating a report indicating the index value calculated in (e).

10. The method of claim 6, further comprising forwarding the report to a medical practitioner.

11. The method of claim 1, wherein the pancreatic cancer is an exocrine pancreatic cancer.

12. The method of claim 11, wherein the pancreatic cancer is pancreatic ductal adenocarcinoma (PDAC).

13. The method of claim 9, wherein at least one hydroxymethylation biomarker exhibits an increase in hydroxymethylation level in the patient relative to the reference hydroxymethylation profile.

14. The method of claim 9, wherein at least one hydroxymethylation biomarker exhibits a decrease in hydroxymethylation level in a patient relative to the reference hydroxymethylation profile.

15. The method of claim 1, wherein the hydroxymethylation biomarkers comprise loci that are associated with one or more of the following genes: ADARB2-AS1, ANKRD36B, ASAH2B, ATG4B, ATP8B1, BOLA1, C11orf88, C17orf97, C1orf170, C3orf36, C8orf74, CAMSAP2, CCDC54, CCDC59, CKAP2, CLK2P, CRTCL1, CSRP2, CYB5D1, DNAJC27, DYNAP, FAM166A, FAM188B, FAM196A, FAM86JP, FAT4, FBXO5, FGF2, FUT2, GAS2L2, GAS6, GGACT, GLRX5, GPX1, GPX5, HBD, HLA-A, HTR1F, IL36G, KANSL1, KCNH6, KCTD15, KLHL38, KLK2, KRT6B, LAMC1, LGALS14, LGALS8-AS1, LIFR, LINC00266-1, LINC00310, LOC100130452, LOC100130557, LOC100130894, LOC100288778, LOC100505633, LOC100505648, LOC100505738, LOC100652909, LOC389033, LOC90784, LRRC37A2, MED11, MRPL23-AS1, NAT8L, NEUROD1, NEUROG2, NME5, NOMO3, NPRL2, NXN, ODF3L1, ODF3L2, OSCP1, PARD6G, PGAM1, PLA2G2E, PLSCR4, PPAP2A, PPP1R15A, PPP1R3E, RASL10B, REXO1L1, RIMBP3, RNF126P1, RNU6-76, RPP25, RPS27, SH3PXD2B, SHISA4, SLC25A38, SLC4A1, SLC05A1, SPDEF, SRSF6, STRA6, SYNM, TBCB, TDRD6, TEX26, TMEM253, TNFSF13B, TTC14, TUBA4A, UBB, VAMP8, VGLL2, WASH2P, WNT9B, XBP1, and ZNF789.

16. The method of claim 1, wherein the hydroxymethylation biomarkers comprise loci that are associated with one or more of the following genes: GATA4, GATA6, PROX1, ONECUT2, YAP1, TEAD1, ONECUT2/ONECUT1-TCGA, IGF1, and IGF2.

17. The method of claim 2, wherein the genetic mutations comprise a mutation in a gene selected from BRCA2, BRCA1, CDKN2A, ATM, STK11, PRSS1, MLH1, PALB2, KRAS, CDKN2A, TP53, SMAD4, and combinations thereof

18. The method of claim 1, wherein step (b) comprises ligating adapters onto the DNA, functionalizing 5hmC residues in the DNA with an affinity tag that allows selective capture of tagged cfDNA, and removing the tagged cfDNA from the sample.

19. The method of claim 18, wherein the affinity tag is comprised of a biotin moiety, and functionalizing the 5hmC residues comprises biotinylation.

20. The method of claim 19, wherein the biotinylation is carried out by covalently attaching a chemoselective group to 5hmC residues and then reacting the chemoselective group with a functionalized biotin moiety.

21. The method of claim 20, wherein the chemoselective group is UDP glucose-6-azide and the functionalized biotin moiety is an alkyne-functionalized biotin, such that the chemoselective group reacts with the functionalized biotin in a click chemistry reaction.

22. The method of claim 19, wherein the affinity tagged cfDNA is captured with a solid support having a surface functionalized with a biotin-binding protein, to provide cfDNA bound to the solid support.

23. The method of claim 22, wherein step (b) further comprises: amplifying the cfDNA without releasing the captured cfDNA from the support, to give a plurality of amplicons; sequencing the amplicons; and quantifying the nucleic acids that map to the reference loci from the sequence reads.

24. The method of claim 23, wherein amplification comprises PCR.

25. The method of claim 18, wherein the adapters additionally comprise at least one unique feature identifier (UFI) sequence.

26. The method of claim 25, wherein the at least one UFI sequence includes a source identifier UFI sequence.

27. The method of claim 25, wherein the at least one UFI sequence is a molecular UFI that enables molecular counting.

28. The method of claim 25, wherein the at least one UFI sequence is a molecular UFI that enables molecular counting.

29. A method for monitoring an identified pancreatic lesion in a patient, the method comprising:

- (a) obtaining an initial cell-free DNA sample from the patient;
- (b) enriching for hydroxymethylated DNA in the initial sample;
- (c) quantifying the nucleic acids in the enriched initial sample that map to each of a plurality of selected loci in a reference hydroxymethylation profile, wherein each selected locus comprises a hydroxymethylation biomarker;
- (d) comparing, at each locus, the hydroxymethylation level of the enriched cell-free DNA in the initial sample with the hydroxymethylation level in the reference profile, to ascertain differences in hydroxymethylation levels between the sample and the reference profile for each biomarker;
- (e) generating an initial hydroxymethylation profile for the patient comprising the hydroxymethylation level of the enriched cell-free DNA in the initial sample, at each locus;
- (f) repeating steps (a) through (c) at a later time with a subsequent cell-free DNA sample obtained from the patient;
- (g) generating a subsequent hydroxymethylation profile for the patient comprising the hydroxymethylation level of the enriched cell-free DNA in the subsequent sample, at each locus; and
- (h) comparing, at each locus, the hydroxymethylation level of the enriched cell-free DNA in the subsequent sample to the hydroxymethylation level of the enriched cell-free DNA in the initial sample, to ascertain a change in the pancreatic lesion.

30. The method of claim 29, wherein steps (f) through (h) are repeated at selected time intervals throughout an extended monitoring period.

31. The method of claim 29, wherein the change in the pancreatic lesion comprises an increase in size.

32. The method of claim 29, wherein the change in the pancreatic lesion comprises a decrease in size.

33. The method of claim 29, wherein the reference hydroxymethylation profile represents a composite of a plurality of hydroxymethylation profiles for individuals who have had a pancreatic lesion identified in an imaging scan.

34. The method of claim 29, wherein the patient has a risk factor for pancreatic cancer and the reference hydroxymethylation profile represents a composite of a plurality of hydroxymethylation profiles for individuals who have the risk factor.

35. The method of claim 34, wherein the risk factor is pancreatitis and the reference hydroxymethylation profile represents a composite of a plurality of hydroxymethylation profiles for individuals who have been diagnosed with pancreatitis.

36. The method of claim 29, wherein the cell-free DNA sample is extracted from a blood sample.

37. The method of claim 29, wherein the cell-free DNA sample is extracted from pancreatic cyst fluid.

38. The method of claim 29, wherein the hydroxymethylation biomarkers comprise loci that are associated with one or more of the following genes: ADARB2-AS1, ANKRD36B, ASAH2B, ATG4B, ATP8B1, BOLA1, C11orf88, C17orf97, C1orf170, C3orf36, C8orf74, CAMSAP2, CCDC54, CCDC59, CKAP2, CLK2P, CRTCL1, CSRP2, CYB5D1, DNAJC27, DYNAP, FAM166A, FAM188B, FAM196A, FAM86JP, FAT4, FBXO5, FGF2, FUT2, GAS2L2, GAS6, GGA2, GLRX5, GPX1, GPX5, HBD, HLA-A, HTR1F, IL36G, KANSL1, KCNH6, KCTD15, KLHL38, KLK2, KRT6B, LAMC1, LGALS14, LGALS8-AS1, LIFR, LINC00266-1, LINC00310, LOC100130452, LOC100130557, LOC100130894, LOC100288778, LOC100505633, LOC100505648, LOC100505738, LOC100652909, LOC389033, LOC90784, LRRC37A2, MED11, MRPL23-AS1, NAT8L, NEUROD1, NEUROG2, NME5, NOMO3, NPRL2, NXN, ODF3L1, ODF3L2, OSCP1, PARD6G, PGAM1, PLA2G2E, PLSCR4, PPAP2A, PPP1R15A, PPP1R3E, RASL10B, REXO1L1, RIMBP3, RNF126P1, RNU6-76, RPP25, RPS27, SH3PXD2B, SHISA4, SLC25A38, SLC4A1, SLC05A1, SPDEF, SRSF6, STRA6,

SYNM, TBCB, TDRD6, TEX26, TMEM253, TNFSF13B, TTC14, TUBA4A, UBB, VAMP8, VGLL2, WASH2P, WNT9B, XBP1, and ZNF789.

39. The method of claim 38, wherein the hydroxymethylation biomarkers additionally comprise loci that are associated with one or more of the following genes: GATA4, GATA6, PROX1, ONECUT2, YAP1, TEAD1, ONECUT2/ONECUT1-TCGA, IGF1, and IGF2.

40. The method of claim 29, wherein step (b) comprises ligating adapters onto the DNA, functionalizing 5hmC residues in the DNA with an affinity tag that allows selective capture of tagged cfDNA, and removing the tagged cfDNA from the sample.

41. The method of claim 40, wherein the affinity tag is comprised of biotin, and functionalizing the 5hmC residues comprises biotinylation.

42. The method of claim 41, wherein the biotinylation is carried out by covalently attaching a chemoselective group to 5hmC residues and then reacting the chemoselective group with a functionalized biotin moiety.

43. The method of claim 42, wherein the chemoselective group is UDP glucose-6-azide and the functionalized biotin moiety is an alkyne-functionalized biotin, such that the chemoselective group reacts with the functionalized biotin in a click chemistry reaction.

44. The method of claim 40, wherein the affinity tagged cfDNA is captured with a solid support having a surface functionalized with a biotin-binding protein, to provide cfDNA bound to the solid support.

45. The method of claim 44, wherein step (b) further comprises: amplifying the cfDNA without releasing the captured cfDNA from the support, to give a plurality of amplicons; sequencing the amplicons; and quantifying the nucleic acids that map to the reference loci from the sequence reads.

46. The method of claim 45, wherein amplification comprises PCR.

47. The method of claim 40, wherein the adapters additionally comprise at least one unique feature identifier (UFI) sequence.

48. The method of claim 47, wherein the at least one UFI sequence includes a source identifier UFI sequence.

49. The method of claim 47, wherein the at least one UFI sequence is a molecular UFI that enables molecular counting.

50. The method of claim 47, wherein the at least one UFI sequence is a molecular UFI that enables molecular counting.

51. A method for managing a patient with a pancreatic lesion identified in an imaging scan, the method comprising:

- (a) obtaining an initial cell-free DNA sample from the patient;
- (b) enriching for hydroxymethylated DNA in the sample;
- (c) quantifying the nucleic acids in the enriched initial sample that map to each of a plurality of selected loci in a reference hydroxymethylation profile, wherein each selected locus comprises a hydroxymethylation biomarker;
- (d) comparing, at each locus, the hydroxymethylation level of the enriched cell-free DNA in the initial sample with the hydroxymethylation level in the reference profile, to ascertain differences in hydroxymethylation levels between the sample and the reference profile for each biomarker;
- (e) generating an initial hydroxymethylation profile for the patient comprising the hydroxymethylation level of the enriched cell-free DNA in the initial sample, at each locus;
- (f) repeating steps (a) through (c) at a later time with a subsequent cell-free DNA sample obtained from the patient;
- (g) generating a subsequent hydroxymethylation profile for the patient comprising the hydroxymethylation level of the enriched cell-free DNA in the subsequent sample, at each locus;
- (h) comparing, at each locus, the hydroxymethylation level of the enriched cell-free DNA in the subsequent sample to the hydroxymethylation level of the enriched cell-free DNA in the initial sample, to ascertain a change in the pancreatic lesion; and

(i) based on the comparison in step (e), determining whether to treat the patient.

52. The method of claim 51, where step (i) comprises determining that treatment is necessary.

53. The method of claim 52, wherein the treatment is selected based on the change in the patient's hydroxymethylation profile at one or more of the selected loci.

54. The method of claim 53, wherein treating the patient comprises radiation therapy, chemotherapy, surgical resection of the lesion, or a combination thereof.

55. The method of claim 51, comprising repeating steps (a) through (h) are repeated at selected time intervals throughout an extended monitoring period.

56. The method of claim 51, wherein the cell-free DNA sample is extracted from a blood sample.

57. The method of claim 51, wherein the cell-free DNA sample is extracted from pancreatic cyst fluid.

58. The method of claim 51, wherein the hydroxymethylation biomarkers comprise loci that are associated with one or more of the following genes: ADARB2-AS1, ANKRD36B, ASAH2B, ATG4B, ATP8B1, BOLA1, C11orf88, C17orf97, C1orf170, C3orf36, C8orf74, CAMSAP2, CCDC54, CCDC59, CKAP2, CLK2P, CRTCL1, CSRP2, CYB5D1, DNAJC27, DYNAP, FAM166A, FAM188B, FAM196A, FAM86JP, FAT4, FBXO5, FGF2, FUT2, GAS2L2, GAS6, GGACT, GLRX5, GPX1, GPX5, HBD, HLA-A, HTR1F, IL36G, KANSL1, KCNH6, KCTD15, KLHL38, KLK2, KRT6B, LAMC1, LGALS14, LGALS8-AS1, LIFR, LINC00266-1, LINC00310, LOC100130452, LOC100130557, LOC100130894, LOC100288778, LOC100505633, LOC100505648, LOC100505738, LOC100652909, LOC389033, LOC90784, LRRC37A2, MED11, MRPL23-AS1, NAT8L, NEUROD1, NEUROG2, NME5, NOMO3, NPRL2, NXN, ODF3L1, ODF3L2, OSCP1, PARD6G, PGAM1, PLA2G2E, PLSCR4, PPAP2A, PPP1R15A, PPP1R3E, RASL10B, REXO1L1, RIMBP3, RNF126P1, RNU6-76, RPP25,

RPS27, SH3PXD2B, SHISA4, SLC25A38, SLC4A1, SLCO5A1, SPDEF, SRSF6, STRA6, SYNM, TBCB, TDRD6, TEX26, TMEM253, TNFSF13B, TTC14, TUBA4A, UBB, VAMP8, VGLL2, WASH2P, WNT9B, XBP1, and ZNF789.

59. The method of claim 58, wherein the hydroxymethylation biomarkers additionally comprise loci that are associated with one or more of the following genes: GATA4, GATA6, PROX1, ONECUT2, YAP1, TEAD1, ONECUT2/ONECUT1-TCGA, IGF1, and IGF2.

60. A method for monitoring the effectiveness of treatment in a patient with a pancreatic lesion identified on an imaging scan, the method comprising:

- (a) obtaining an initial cell-free DNA sample from a patient who is being treated;
- (b) enriching for hydroxymethylated DNA in the sample;
- (c) quantifying the nucleic acids in the enriched initial sample that map to each of a plurality of selected loci in a reference hydroxymethylation profile, wherein each selected locus comprises a hydroxymethylation biomarker;
- (d) comparing, at each locus, the hydroxymethylation level of the enriched cell-free DNA in the initial sample with the hydroxymethylation level in the reference profile, to ascertain differences in hydroxymethylation levels between the sample and the reference profile for each biomarker;
- (e) generating an initial hydroxymethylation profile for the patient comprising the hydroxymethylation level of the enriched cell-free DNA in the initial sample, at each locus;
- (f) repeating steps (a) through (c) at a later time with a subsequent cell-free DNA sample obtained from the patient;
- (g) generating a subsequent hydroxymethylation profile for the patient comprising the hydroxymethylation level of the enriched cell-free DNA in the subsequent sample, at each locus;
- (h) comparing, at each locus, the hydroxymethylation level of the enriched cell-free DNA in the subsequent sample to the hydroxymethylation level of the enriched cell-free DNA in the initial sample, to ascertain a change in the pancreatic lesion; and
- (i) if the comparison in step (e) evidences changes in the patient's hydroxymethylation profile that correlate with a progression toward cancer, changing the treatment protocol.

61. The method of claim 60, wherein the cell-free DNA sample is extracted from a blood sample.

62. The method of claim 60, wherein the cell-free DNA sample is extracted from pancreatic cyst fluid.

63. The method of claim 60, wherein the hydroxymethylation biomarkers comprise loci that are associated with one or more of the following genes: ADARB2-AS1, ANKRD36B, ASAH2B, ATG4B, ATP8B1, BOLA1, C11orf88, C17orf97, C1orf170, C3orf36, C8orf74, CAMSAP2, CCDC54, CCDC59, CKAP2, CLK2P, CRTCL1, CSRP2, CYB5D1, DNAJC27, DYNAP, FAM166A, FAM188B, FAM196A, FAM86JP, FAT4, FBXO5, FGF2, FUT2, GAS2L2, GAS6, GGACT, GLRX5, GPX1, GPX5, HBD, HLA-A, HTR1F, IL36G, KANSL1, KCNH6, KCTD15, KLHL38, KLK2, KRT6B, LAMC1, LGALS14, LGALS8-AS1, LIFR, LINC00266-1, LINC00310, LOC100130452, LOC100130557, LOC100130894, LOC100288778, LOC100505633, LOC100505648, LOC100505738, LOC100652909, LOC389033, LOC90784, LRRC37A2, MED11, MRPL23-AS1, NAT8L, NEUROD1, NEUROG2, NME5, NOMO3, NPRL2, NXN, ODF3L1, ODF3L2, OSCP1, PARD6G, PGAM1, PLA2G2E, PLSCR4, PPAP2A, PPP1R15A, PPP1R3E, RASL10B, REXO1L1, RIMBP3, RNF126P1, RNU6-76, RPP25, RPS27, SH3PXD2B, SHISA4, SLC25A38, SLC4A1, SLCO5A1, SPDEF, SRSF6, STRA6, SYNM, TBCB, TDRD6, TEX26, TMEM253, TNFSF13B, TTC14, TUBA4A, UBB, VAMP8, VGLL2, WASH2P, WNT9B, XBP1, and ZNF789.

64. The method of claim 63, wherein the hydroxymethylation biomarkers additionally comprise loci that are associated with one or more of the following genes: GATA4, GATA6, PROX1, ONECUT2, YAP1, TEAD1, ONECUT2/ONECUT1-TCGA, IGF1, and IGF2.

65. The method of claim 60, wherein the pancreatic cancer is an exocrine pancreatic cancer.

66. The method of claim 65, wherein the pancreatic cancer is PDAC.

67. A method for reducing the risk that a pancreatic lesion surgically removed from a patient is benign, comprising, prior to surgery,

- (a) obtaining a cell-free DNA sample from the patient;
- (b) enriching for hydroxymethylated DNA in the sample;
- (c) quantifying the nucleic acids in the enriched sample that map to each of a plurality of selected loci in a reference hydroxymethylation profile, wherein each selected locus comprises a hydroxymethylation biomarker;
- (d) comparing, at each locus, the hydroxymethylation level of the sample with the hydroxymethylation level in the reference profile, to ascertain differences in hydroxymethylation levels between the sample and the reference profile for each biomarker; and
- (e) calculating an index value representing the risk that the pancreatic lesion is cancerous from the comparison in step (d) combined with at least one additional parameter correlated with the risk that an individual has pancreatic cancer; and
- (f) carrying out surgical resection of the pancreatic lesion only if the index value is greater than a value corresponding to a low risk of cancer.

68. The method of claim 67, wherein the cell-free DNA sample is extracted from a blood sample.

69. The method of claim 67, wherein the cell-free DNA sample is extracted from pancreatic cyst fluid.

70. The method of claim 67, wherein the hydroxymethylation biomarkers comprise loci that are associated with one or more of the following genes: ADARB2-AS1, ANKRD36B, ASAH2B, ATG4B, ATP8B1, BOLA1, C11orf88, C17orf97, C1orf170, C3orf36, C8orf74, CAMSAP2, CCDC54, CCDC59, CKAP2, CLK2P, CRTCL1, CSRP2, CYB5D1, DNAJC27, DYNAP, FAM166A, FAM188B, FAM196A, FAM86JP, FAT4, FBXO5, FGF2, FUT2, GAS2L2, GAS6, GGACT, GLRX5, GPX1, GPX5, HBD, HLA-A, HTR1F, IL36G, KANSL1, KCNH6, KCTD15, KLHL38, KLK2, KRT6B, LAMC1, LGALS14, LGALS8-AS1, LIFR, LINC00266-1, LINC00310, LOC100130452, LOC100130557, LOC100130894, LOC100288778, LOC100505633, LOC100505648, LOC100505738, LOC100652909, LOC389033, LOC90784, LRRC37A2, MED11,

MRPL23-AS1, NAT8L, NEUROD1, NEUROG2, NME5, NOMO3, NPRL2, NXN, ODF3L1, ODF3L2, OSCP1, PARD6G, PGAM1, PLA2G2E, PLSCR4, PPAP2A, PPP1R15A, PPP1R3E, RASL10B, REXO1L1, RIMBP3, RNF126P1, RNU6-76, RPP25, RPS27, SH3PXD2B, SHISA4, SLC25A38, SLC4A1, SLCO5A1, SPDEF, SRSF6, STRA6, SYNM, TBCB, TDRD6, TEX26, TMEM253, TNFSF13B, TTC14, TUBA4A, UBB, VAMP8, VGLL2, WASH2P, WNT9B, XBP1, and ZNF789.

71. The method of claim 70, wherein the hydroxymethylation biomarkers additionally comprise loci that are associated with one or more of the following genes: GATA4, GATA6, PROX1, ONECUT2, YAP1, TEAD1, ONECUT2/ONECUT1-TCGA, IGF1, and IGF2.

72. A method for evaluating the risk that a subject will develop pancreatic cancer, comprising:

- (a) obtaining a cell-free DNA sample from the subject;
- (b) enriching for hydroxymethylated DNA in the sample;
- (c) quantifying the nucleic acids in the enriched sample that map to each of a plurality of selected loci in a reference hydroxymethylation profile, wherein each selected locus comprises a hydroxymethylation biomarker;
- (d) comparing, at each locus, the hydroxymethylation level of the sample with the hydroxymethylation level in the reference profile, to ascertain differences in hydroxymethylation levels between the sample and the reference profile for each biomarker; and
- (e) calculating an index value representing the subject's risk of developing pancreatic cancer from the comparison in step (d) combined with at least one additional parameter correlated with the risk that an individual will develop pancreatic cancer.

73. A kit for carrying out the method of claim 1, comprising:
at least one reagent for the determination of hydroxymethylation level at each of a plurality of hydroxymethylation biomarker loci in a cell-free DNA sample;
a solid support for capturing affinity-tagged 5hmC-containing cell-free DNA in the sample; and
written instructions for the use of the at least one reagent and the solid support in carrying out the method.

74. The kit of claim 73, further including instructions for accessing and using software designed to perform modeling and prediction.

75. A kit for carrying out the method of claim 1, comprising:
a DNA β -glucosyl transferase;
UDP glucose modified with a chemoselective group;
a biotin moiety;
a solid support having a surface functionalized with a biotin-binding protein;
an adaptor comprising a molecular barcode; and
written instructions for carrying out the method.

76. The kit of claim 75, further including instructions for accessing and using software designed to perform modeling and prediction.

77. A method for determining the likelihood that an individual at risk for developing pancreatic cancer has pancreatic cancer, the method comprising:
(a) obtaining a cell-free DNA sample from the patient;
(b) enriching for hydroxymethylated DNA in the sample;
(c) quantifying the nucleic acids in the enriched sample that map to each of a plurality of selected loci in a reference hydroxymethylation profile, wherein each selected locus comprises a hydroxymethylation biomarker;
(d) comparing, at each locus, the hydroxymethylation level of the sample with the hydroxymethylation level in the reference profile, to ascertain differences in hydroxymethylation levels between the sample and the reference profile for each biomarker;
and
(e) calculating an index value representing the likelihood that the individual has pancreatic cancer from the comparison in step (d).

78. The method of claim 77, further including, prior to step (a), identifying the individual as being at risk for developing pancreatic cancer from one or more parameters selected from: an identified pancreatic lesion; pancreatic inflammation; jaundice; age; weight; gender; ethnicity; family history; genetic mutations; diabetes; physical activity; diet; pro-inflammatory cytokine levels; and cigarette smoking.

79. The method of claim 77, further including determining the likelihood that the individual has at least one additional type of cancer.

80. The method of claim 79, wherein the at least one additional type of cancer is selected from: bladder cancer; cancers of the blood and bone marrow; brain cancer; breast cancer; cervical cancer; colorectal cancer; esophageal cancer; liver cancer; lung cancer; ovarian cancer; prostate cancer; renal cancer; skin cancer; testicular cancer; thyroid cancer; and uterine cancer.

81. In a multi-cancer test that determines the likelihood that an individual has pancreatic cancer and at least one additional type of cancer selected from colorectal, esophageal, lung, and liver cancer, the improvement which comprises determining the likelihood that the individual has pancreatic cancer by:

- (a) obtaining a cell-free DNA sample from the individual;
- (b) enriching for hydroxymethylated DNA in the sample;
- (c) quantifying the nucleic acids in the enriched sample that map to each of a plurality of selected loci in a reference hydroxymethylation profile, wherein each selected locus comprises a hydroxymethylation biomarker;
- (d) comparing, at each locus, the hydroxymethylation level of the sample with the hydroxymethylation level in the reference profile, to ascertain differences in hydroxymethylation levels between the sample and the reference profile for each biomarker; and
- (e) calculating an index value representing the likelihood that the individual has pancreatic cancer from the comparison in step (d).

82. The improved multi-cancer test of claim 81, further including eliminating false positives for the at least one additional type of cancer prior to (a).

83. The improved multi-cancer test of claim 81, further including eliminating false negatives for the at least one additional type of cancer prior to (a).

84. A kit for carrying out the method of claim 77, comprising:

at least one reagent for the determination of hydroxymethylation level at each of a plurality of hydroxymethylation biomarker loci in a cell-free DNA sample;
a solid support for capturing affinity-tagged 5hmC-containing cell-free DNA in the sample; and
written instructions for the use of the at least one reagent and the solid support in carrying out the method.

85. The kit of claim 84, further including instructions for accessing and using software designed to perform modeling and prediction.

86. A kit for carrying out the method of claim 77, comprising:
a DNA β -glucosyl transferase;
UDP glucose modified with a chemoselective group;
a biotin moiety;
a solid support having a surface functionalized with a biotin-binding protein;
an adaptor comprising a molecular barcode; and
written instructions for carrying out the method.

87. The kit of claim 86, further including instructions for accessing and using software designed to perform modeling and prediction.