DNA METHYLATION AND GENE EXPRESSION AS DETERMINANTS OF GENOME-WIDE CELL-FREE DNA FRAGMENTATION

The present application claims the benefit of priority of U.S. provisional application no. 63/521,666 filed June 17, 2023, which is incorporated by reference herein in its entirety.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH

[0001] This invention was made with government support under grants CA006973, CA062924, CA121113 and CA233259 awarded by the National Institutes of Health. The government has certain rights in the invention.

FIELD

[0002] This disclosure relates to methods and compositions for assessing and/or treating mammals (e.g., humans) having cancer. In particular, methods relate to analyzing epigenetic changes and cfDNA fragmentation for non-invasive disease detection.

BACKGROUND

[0003] Cell-free DNA (cfDNA) has been the focus of research in blood-based biomarkers for early detection and monitoring of cancer. Normally, cellular DNA is packaged and condensed within chromosomes by wrapping around histone cores[1,2]. In the process of and after cellular death, DNA is digested by DNAses, in part to prevent the release of unbound DNA which can act as auto- antigens[3–6]. DNA-fragments that are tightly wrapped around histone cores, collectively called nucleosomes, appear to be protected from further digestion[7]. These fragments are those that are typically represented in cfDNA and can be collected with a simple blood draw. The characteristics and origins of cfDNA fragmentation in the blood are poorly understood.

SUMMARY

[0004] Methods for the diagnosing and treatment of cancer include detecting methylation and expression differences affecting cell free (cfDNA) size and coverage in patients.

[0005] In certain aspects, a method for determining circulating cell-free DNA (cfDNA) fragmentation in a sample in provided, comprising assaying genomic sequences to identify

cfDNA end positions; analyzing frequency of cfDNA breakpoints at a plurality of positions in the genomic sequences; thereby, determining circulating cell-free DNA (cfDNA) fragmentation.

[0006] In certain aspects, a method for determining circulating cell-free DNA (cfDNA) fragmentation in a sample is provided, the comprising a) assaying cfDNA sequences to identify the genomic location and end positions of cfDNA fragments; b) analyzing frequency of cfDNA breakpoints at a plurality of positions in the genomic sequences; thereby determining circulating cell-free DNA (cfDNA) fragmentation.

[0007] In certain embodiments of the above methods, the genomic sequences are assayed by whole genome sequencing or obtaining whole genome sequences from a database and pooling cfDNA sequences.

[0008] In certain embodiments, analysis of the frequency of cfDNA breakpoints comprises calculating a ratio of number of cfDNA fragments starting or ending at a particular position compared to the number of fragments with start or end positions within 50 bp surrounding that particular position.

[0009] In certain embodiments of the above methods, analysis of the frequency of cfDNA breakpoints comprises calculating the ratios of number of cfDNA fragments starting or the ratio of number of cfDNA fragments ending at a particular position compared to the number of fragments with start or end positions within 50 bp surrounding that particular position.

[0010] In certain embodiments, cfDNA fragments comprising similar end-position sequences comprise similar motifs. In certain embodiments, the motifs comprise a thymine or an adenine before a start of the cfDNA fragment sequence and two cytosines (A/T|CC) or a cytosine followed by a guanine (A/T|CG) as first two nucleotides of the cfDNA fragment sequence.

[0011] In certain embodiments, the motifs comprise a thymine or an adenine before a start of the cfDNA fragment sequence and two cytosines (A/T|CC) or a cytosine followed by a guanine (A/T|CG) as first two nucleotides of the cfDNA fragment sequence or where these sequences are the inverse complement at the end of a cfDNA fragment

[0012] In certain embodiments, the frequency of the motifs is increased in healthy subjects as compared to a subject having cancer. For example, in such embodiments, , the frequency of the motifs is increased in healthy subjects by at least 0.5, 1. 2, 3, 4, 5, 6, 7, 8, 9, 10,

12, 14, 16, 18, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100 150 or 200 percent relative to the frequency of the motifs in a subject having cancer. In certain embodiments, the frequency of the motifs is decreased in healthy subjects as compared to a subject having cancer. For example, in such embodiments, , the frequency of the motifs is decreased in healthy subjects by at least 0.5, 1. 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100 150 or 200 percent relative to the frequency of the motifs in a subject having cancer.

[0013] In certain embodiments, the frequency of A/T|CC motifs is greater than the frequency of A/T|CG motifs. In certain embodiments, the A/T|CG motifs are positioned close to a histone H1 linker or centered between 100 base pairs to 200 base pairs from the histone H1 linker. In certain embodiments, interior regions of cfDNA fragments are enriched with adenines and thymines.

[0014] In certain embodiments, the method further comprises mapping cfDNA fragments to the genome and comparing the cfDNA fragment end sequences to methylated and unmethylated CpG sites of cfDNA from healthy subjects.

[0015] In certain embodiments, methylated CpGs are enriched at the ends of A/T|CG cfDNA fragment sequence.

[0016] In certain embodiments, quantitative assessment of enrichment of cfDNA fragment ends at CpGs comprise calculating for each CpG a fraction of cfDNA fragments starting or ending at CpG dinucleotide positions over number of cfDNA fragments with start or end positions within 50 bp around each CpG.

[0017] In certain embodiments, cfDNA fragments obtained from an X chromosome among healthy subjects comprise increased cfDNA fragments ending with CG and cfDNA fragment ends with CCG are decreased at locations of X chromosome CpG islands in female subjects as compared to male subjects. In certain embodiments, cfDNA sequence coverage of enriched cfDNA fragment end sequences comprise increased methylation across regions of methylated CpG islands as compared to reduced or lower occurrence cfDNA fragment end sequences.

[0018] In certain embodiments, gene expression at transcription start sites (TSS) is inversely related to cfDNA coverage at TSS.

[0019] The subject e.g. mammal particularly a human can have previously been administered a cancer treatment to treat the cancer. The method may further comprise administering a cancer treatment after determining circulating cell-free DNA (cfDNA) fragmentation, for example an administered cancer treatment can be surgery, adjuvant chemotherapy, neoadjuvant chemotherapy, radiation therapy, hormone therapy, cytotoxic therapy, immunotherapy, adoptive T cell therapy, targeted therapy, or combinations thereof. The method also can include monitoring the mammal for the presence of cancer after administration of the cancer treatment and/or after determining circulating cell-free DNA (cfDNA) fragmentation.

[0020] In another aspect, a method of diagnosing cancer and treating a subject is providing and suitably comprises obtaining a sample from a subject, assaying for changes in circulating cell-free DNA (cfDNA) fragment sizes as compared to normal and tumor-derived cfDNA controls, assaying for CpG methylation on cfDNA fragment ends and assessing fragment end representation at CG and CCG sites through low coverage whole genome cfDNA analyses; diagnosing the subject with cancer; and, treating the subject with one or more chemotherapeutic agents, radiotherapy, surgery and combinations thereof. In certain embodiments, an increase in cfDNA fragments ending in N|CCG as compared to a normal control is diagnostic of cancer. In certain embodiments, a decrease in cfDNA fragments ending with N|CCG is indicative of a subject without cancer. In certain embodiments, an increase of cfDNA fragments ending with CG as compared to normal controls is diagnostic of cancer. In certain embodiments, the cfDNA fragments are obtained from genomic regions with increased methylation as compared to normal genomic methylation controls. In certain embodiments, further comprises incorporating distribution of fragment end positions at CG and CCG sites in a gradient boosted tree machine learning model.

[0021] In another aspect, a method for diagnosing and treating a subject diagnosed with cancer, comprises assaying genomic sequences to identify genome-wide CpG methylation; analyzing frequency of cfDNA breakpoints at a plurality of positions in the genomic sequences and determining circulating cell-free DNA (cfDNA) fragmentation, wherein, recurrent cfDNA fragment end enrichment at CpG sites correlate with higher genome-wide methylation levels and smaller cfDNA fragments diagnostic of cancer; diagnosing the subject with cancer if

hypomethylation and/or increased gene expression or a decrease in cfDNA fragment size is detected; and, treating the subject with one or more chemotherapies, radiation, surgery or combinations thereof. In certain embodiments, the genomic sequences are assayed by whole genome sequencing or obtaining whole genome sequences from a database and pooling cfDNA sequences.

[0022] In certain embodiments, analysis of the frequency of cfDNA breakpoints comprises calculating a ratio of number of cfDNA fragments starting or ending at a particular position compared to the number of fragments with start or end positions within 50 bp surrounding that particular position.

[0023] In certain embodiments, cfDNA fragments comprising similar end-position sequences comprise similar motifs. In certain embodiments, the motifs comprise a thymine or an adenine before a start of the cfDNA fragment sequence and two cytosines (A/T|CC) or a cytosine followed by a guanine (A/T|CG) as first two nucleotides of the cfDNA fragment sequence.

[0024] In certain embodiments, the frequency of the motifs is increased in healthy subjects as compared to a subject having cancer. In certain embodiments, the frequency of A/T|CC motifs is greater than the frequency of A/T|CG motifs. In certain embodiments, the A/T|CG motifs are positioned close to a histone H1 linker or centered between 100 base pairs to 200 base pairs from the histone H1 linker. In certain embodiments, interior regions of cfDNA fragments are enriched with adenines and thymines. In certain embodiments, the method further comprises mapping cfDNA fragments to the genome and comparing the cfDNA fragment end sequences to methylated and unmethylated CpG sites of cfDNA from healthy subjects. In certain embodiments, methylated CpGs are enriched at the ends of A/T|CG cfDNA fragment sequence. In certain embodiments, quantitative assessment of enrichment of cfDNA fragment ends at CpGs comprise calculating for each CpG a fraction of cfDNA fragments starting or ending at CpG dinucleotide positions over number of cfDNA fragments with start or end positions within 50 bp around each CpG.

[0025] In certain embodiments, cfDNA fragments obtained from an X chromosome among healthy subjects comprise increased cfDNA fragments ending with CG and cfDNA

fragment ends with CCG are decreased at locations of X chromosome CpG islands in female subjects as compared to male subjects.

[0026] In certain embodiments, cfDNA sequence coverage of enriched cfDNA fragment end sequences comprise increased methylation across regions of methylated CpG islands as compared to reduced or lower occurrence cfDNA fragment end sequences. In certain embodiments, gene expression at transcription start sites (TSS) is inversely related to cfDNA coverage at TSS. The subject e.g. mammal particularly a human can have previously been administered a cancer treatment to treat the cancer. The cancer treatment can be surgery, adjuvant chemotherapy, neoadjuvant chemotherapy, radiation therapy, hormone therapy, cytotoxic therapy, immunotherapy, adoptive T cell therapy, targeted therapy, or combinations thereof. The method also can include monitoring the mammal for the presence of cancer after administration of the cancer treatment.

[0027] In another aspect, a method for determining a cfDNA fragmentation profile of a mammal comprises processing cfDNA fragments obtained from a sample obtained from the mammal into sequencing libraries, subjecting the sequencing libraries to whole genome sequencing (e.g., low-coverage whole genome sequencing) to obtain sequenced fragments, mapping the sequenced fragments to a genome to obtain windows of mapped sequences, and analyzing the windows of mapped sequences to determine cfDNA fragment lengths. The mapped sequences can include tens to thousands of windows. The windows of mapped sequences can be non-overlapping windows. The windows of mapped sequences can each include up to about 0.25, 0.5, 1, 2, 3, 4, 5 or 6 million or more base pairs. The cfDNA fragmentation profile can be determined within each window. The cfDNA fragmentation profile can include a median fragment size. The cfDNA fragmentation profile can include a fragment size distribution. The cfDNA fragmentation profile can include a ratio of small cfDNA fragments to large cfDNA fragments in the windows of mapped sequences. The cfDNA fragmentation profile can be over the whole genome. The cfDNA fragmentation profile can be over a subgenomic interval (e.g., an interval in a portion of a chromosome). In certain embodiments, the reference cfDNA fragmentation profile can be a cfDNA fragmentation profile of a healthy mammal. In certain embodiments, the reference cfDNA fragmentation profile can

be generated by determining a cfDNA fragmentation profile in a sample obtained from a healthy mammal. In certain embodiments, the reference DNA fragmentation pattern can be a reference nucleosome cfDNA fragmentation profile. In certain embodiments, the cfDNA fragmentation profile can include a median fragment size, where a median fragment size of the cfDNA fragmentation profile is shorter than a median fragment size of the reference cfDNA fragmentation profile. In certain embodiments, the cfDNA fragmentation profile can include a fragment size distribution, where a fragment size distribution of the cfDNA fragmentation profile differs by at least 4, 6, 8, 10, 12, 14, 16, 18 or 20 nucleotides as compared to a fragment size distribution of the reference cfDNA fragmentation profile.

[0028] In certain embodiments, the cfDNA fragmentation profile can include a ratio of small cfDNA fragments to large cfDNA fragments in the windows of mapped sequences, where a small cfDNA fragment is up to 40, 60, 80 or 100 bp to 110, 120, 130, 140 or 150 bp in length, where a large cfDNA fragments is 151 bp to 220 bp in length, and where a correlation of fragment ratios in the cfDNA fragmentation profile is lower than a correlation of fragment ratios of the reference cfDNA fragmentation profile. In certain embodiments, the cfDNA fragmentation profile can include a ratio of small cfDNA fragments to large cfDNA fragments in the windows of mapped sequences, where a small cfDNA fragment is up to 100 bp to 150 bp in length, where a large cfDNA fragments is 151 bp to 220 bp in length, and where a correlation of fragment ratios in the cfDNA fragmentation profile is lower than a correlation of fragment ratios of the reference cfDNA fragmentation profile.

[0029] In certain embodiments, the cfDNA fragmentation profile can include the sequence coverage of small cfDNA fragments in windows across the genome. In certain embodiments, the cfDNA fragmentation profile can include the sequence coverage of large cfDNA fragments in windows across the genome. In certain embodiments, the cfDNA fragmentation profile can include the sequence coverage of small and large cfDNA fragments in windows across the genome. The step of comparing can include comparing the cfDNA fragmentation profile to a reference cfDNA fragmentation profile over the whole genome. The step of comparing can include comparing the cfDNA fragmentation profile to a reference cfDNA

fragmentation profile over a subgenomic interval. The mammal can have previously been administered a cancer treatment to treat the cancer.

[0030] In another aspect, a method for diagnosing and treating a subject diagnosed with cancer is provided, comprising: a) assaying genomic sequences to identify genome-wide CpG methylation; b) analyzing frequency of cfDNA breakpoints at a plurality of positions in the genomic sequences and determining circulating cell-free DNA (cfDNA) fragmentation, wherein, recurrent cfDNA fragment end enrichment at CpG sites correlate with higher genome-wide methylation levels and smaller cfDNA fragments diagnostic of cancer; c) diagnosing the subject with cancer if hypomethylation and/or increased gene expression or a decrease in cfDNA fragment size is detected; and, d) treating the subject with one or more chemotherapies, radiation, surgery or combinations thereof.

[0031] In certain embodiments, the genomic sequences are assayed by whole genome sequencing or obtaining whole genome sequences from a database and pooling cfDNA sequences. In certain embodiments, analysis of the frequency of cfDNA breakpoints comprises calculating a ratio of number of cfDNA fragments starting or ending at a particular position compared to the number of fragments with start or end positions within 50 bp surrounding that particular position. In certain embodiments, cfDNA fragments comprising similar end-position sequences comprise similar motifs. In certain embodiments, the motifs comprise a thymine or an adenine before a start of the cfDNA fragment sequence and two cytosines (A/T|CC) or a cytosine followed by a guanine (A/T|CG) as first two nucleotides of the cfDNA fragment sequence. In certain embodiments, wherein the frequency of the motifs is increased in healthy subjects as compared to a subject having cancer. In certain embodiments, the frequency of A/T|CC motifs is greater than the frequency of A/T|CG motifs. In certain embodiments, the A/T|CG motifs are positioned close to a histone H1 linker or centered between 100 base pairs to 200 base pairs from the histone H1 linker. In certain embodiments, interior regions of cfDNA fragments are enriched with adenines and thymines. In certain embodiments, the methods may further comprise mapping cfDNA fragments to the genome and comparing the cfDNA fragment end sequences to methylated and unmethylated CpG sites of cfDNA from healthy subjects. In certain embodiments, methylated CpGs are enriched at the ends of A/T|CG cfDNA fragment sequence.

In certain embodiments, quantitative assessment of enrichment of cfDNA fragment ends at CpGs comprise calculating for each CpG a fraction of cfDNA fragments starting or ending at CpG dinucleotide positions over number of cfDNA fragments with start or end positions within 50 bp around each CpG. In certain embodiments, cfDNA fragments obtained from an X chromosome among healthy subjects comprise increased cfDNA fragments ending with CG and cfDNA fragment ends with CCG are decreased at locations of X chromosome CpG islands in female subjects as compared to male subjects. In certain embodiments, cfDNA sequence coverage of enriched cfDNA fragment end sequences comprise increased methylation across regions of methylated CpG islands as compared to reduced or lower occurrence cfDNA fragment end sequences. In certain embodiments, gene expression at transcription start sites (TSS) is inversely related to cfDNA coverage at TSS.

[0032] *Definitions*

[0033] Unless otherwise defined, all terms (including technical and scientific terms) used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. It will be further understood that terms, such as those defined in commonly used dictionaries, should be interpreted as having a meaning that is consistent with their meaning in the context of the relevant art and will not be interpreted in an idealized or overly formal sense unless expressly so defined herein.

[0034] As used herein, the singular forms "a", "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. Furthermore, to the extent that the terms "including", "includes", "having", "has", "with", or variants thereof are used in either the detailed description and/or the claims, such terms are intended to be inclusive in a manner similar to the term "comprising."

[0035] The term "about" or "approximately" means within an acceptable error range for the particular value as determined by one of ordinary skill in the art, which will depend in part on how the value is measured or determined, i.e., the limitations of the measurement system. For example, "about" can mean within 1 or more than 1 standard deviation, per the practice in the art. Alternatively, "about" can mean a range of up to 20%, up to 10%, up to 5%, or up to 1% of a given value or range. Alternatively, particularly with respect to biological systems or processes,

the term can mean within an order of magnitude within 5-fold, and also within 2-fold, of a value. Where particular values are described in the application and claims, unless otherwise stated the term "about" meaning within an acceptable error range for the particular value should be assumed.

[0036] The terms "aligned", "alignment", "mapped" or "aligning", "mapping" refer to one or more sequences that are identified as a match in terms of the order of their nucleic acid molecules to a known sequence from a reference genome. Such alignment can be done manually or by a computer algorithm, examples including the Efficient Local Alignment of Nucleotide Data (ELAND) computer program distributed as part of the Illumina Genomics Analysts pipeline. The matching of a sequence read in aligning can be a 100% sequence match or less than 100% (non-perfect match).

[0037] The term "cancer" as used herein is meant, a disease, condition, trait, genotype or phenotype characterized by unregulated cell growth or replication as is known in the art; including liver cancer (including hepatocellular carcinoma (HCC)), lung cancer (including non-small cell lung carcinoma), gastric cancer, colorectal cancer, as well as, for example, leukemias, e.g., acute myelogenous leukemia (AML), chronic myelogenous leukemia (CML), acute lymphocytic leukemia (ALL), and chronic lymphocytic leukemia, AIDS related cancers such as Kaposi's sarcoma; breast cancers; bone cancers such as Osteosarcoma, Chondrosarcomas, Ewing's sarcoma, Fibrosarcomas, Giant cell tumors, Adamantinomas, and Chordomas; Brain cancers such as Meningiomas, Glioblastomas, Lower- Grade Astrocytomas, Oligodendrocytomas, Pituitary Tumors, Schwannomas, and Metastatic brain cancers; cancers of the head and neck including various lymphomas such as mantle cell lymphoma, non-Hodgkins lymphoma, adenoma, squamous cell carcinoma, laryngeal carcinoma, gallbladder and bile duct cancers, cancers of the retina such as retinoblastoma, cancers of the esophagus, gastric cancers, multiple myeloma, ovarian cancer, uterine cancer, thyroid cancer, testicular cancer, endometrial cancer, melanoma, bladder cancer, prostate cancer, pancreatic cancer, sarcomas, Wilms' tumor, cervical cancer, head and neck cancer, skin cancers, nasopharyngeal carcinoma, liposarcoma, epithelial carcinoma, renal cell carcinoma, gallbladder adeno carcinoma, parotid

adenocarcinoma, endometrial sarcoma, multidrug resistant cancers; and proliferative diseases and conditions, such as neovascularization associated with tumor angiogenesis.

[0038] The term "cell free nucleic acid," "cell free DNA," or "cfDNA" refers to nucleic acid fragments that circulate in an individual's body (e.g., bloodstream) and originate from one or more healthy cells and/or from one or more cancer cells. Additionally, cfDNA may come from other sources such as viruses, fetuses, etc.

[0039] The term "cfDNA sequence coverage" refers to the average number of cfDNA molecules overlapping a specific position.

[0040] The term "circulating tumor DNA" or "ctDNA" refers to nucleic acid fragments that originate from tumor cells or other types of cancer cells, which may be released into an individual's bloodstream as result of biological processes such as apoptosis or necrosis of dying cells or actively released by viable tumor cells.

[0041] As used herein, the terms "comprising," "comprise" or "comprised," and variations thereof, in reference to defined or described elements of an item, composition, apparatus, method, process, system, etc. are meant to be inclusive or open ended, permitting additional elements, thereby indicating that the defined or described item, composition, apparatus, method, process, system, etc. includes those specified elements--or, as appropriate, equivalents thereof--and that other elements can be included and still fall within the scope/definition of the defined item, composition, apparatus, method, process, system, etc.

[0042] "Diagnostic" or "diagnosed" means identifying the presence or nature of a pathologic condition. Diagnostic methods differ in their sensitivity and specificity. The "sensitivity" of a diagnostic assay is the percentage of diseased individuals who test positive (percent of "true positives"). Diseased individuals not detected by the assay are "false negatives." Subjects who are not diseased and who test negative in the assay, are termed "true negatives." The "specificity" of a diagnostic assay is 1 minus the false positive rate, where the "false positive" rate is defined as the proportion of those without the disease who test positive. While a particular diagnostic method may not provide a definitive diagnosis of a condition, it suffices if the method provides a positive indication that aids in diagnosis.

[0043] An "effective amount" as used herein, means an amount which provides a therapeutic or prophylactic benefit.

[0044] As used herein, the terms "fragmentation profile," "position dependent differences in fragmentation patterns," and "differences in fragment size and coverage in a position dependent manner across the genome" are equivalent and can be used interchangeably. In some embodiments, determining a cfDNA fragmentation profile in a mammal can be used for identifying a mammal as having cancer. For example, cfDNA fragments obtained from a mammal (e.g., from a sample obtained from a mammal) can be subjected to low coverage whole-genome sequencing, and the sequenced fragments can be mapped to the genome (e.g., in non-overlapping windows) and assessed to determine a cfDNA fragmentation profile. As described herein, a cfDNA fragmentation profile of a mammal having cancer is more heterogeneous (e.g., in fragment lengths) than a cfDNA fragmentation profile of a healthy mammal (e.g., a mammal not having cancer). As such, this disclosure also provides methods and materials for assessing, monitoring, and/or treating mammals (e.g., humans) having, or suspected of having, cancer. In some embodiments, this document provides methods and materials for identifying a mammal as having cancer. For example, a sample (e.g., a blood sample) obtained from a mammal can be assessed to determine the presence and, optionally, the tissue of origin of the cancer in the mammal based, at least in part, on the cfDNA fragmentation profile of the mammal. In some embodiments, methods and materials for monitoring a mammal as having cancer are provided. For example, a sample (e.g., a blood sample) obtained from a mammal can be assessed to determine the presence of the cancer in the mammal based, at least in part, on the cfDNA fragmentation profile of the mammal. In some embodiments, methods and materials for identifying a mammal as having cancer and administering one or more cancer treatments to the mammal to treat the mammal are provided. For example, a sample (e.g., a blood sample) obtained from a mammal can be assessed to determine if the mammal has cancer based, at least in part, on the cfDNA fragmentation profile of the mammal, and one or more cancer treatments can be administered to the mammal.

[0045] The term "genomic nucleic acid," or "genomic DNA," refers to nucleic acid including chromosomal DNA that originates from one or more healthy (e.g., non-tumor) cells or

tumor cells. In various embodiments, genomic DNA can be extracted from a cell derived from a blood cell lineage, such as a white blood cell (WBC).

[0046] "Optional" or "optionally" means that the subsequently described event or circumstance can or cannot occur, and that the description includes instances where the event or circumstance occurs and instances where it does not.

[0047] As used in this specification and the appended claims, the term "or" is generally employed in its sense including "and/or" unless the content clearly dictates otherwise.

[0048] "Parenteral" administration of an immunogenic composition includes, e.g., subcutaneous (s.c.), intravenous (i.v.), intramuscular (i.m.), or intrasternal injection, or infusion techniques.

[0049] The terms "patient" or "individual" or "subject" are used interchangeably herein, and refers to a mammalian subject to be treated, with human patients being preferred. In some embodiments, the methods of the invention find use in experimental animals, in veterinary application, and in the development of animal models for disease, including, but not limited to, rodents including mice, rats, and hamsters, and primates.

[0050] The term "reference genome" as used herein may refer to a digital or previously identified nucleic acid sequence database, assembled as a representative example of a species or subject. Reference genomes may be assembled from the nucleic acid sequences from multiple subjects, sample or organisms and does not necessarily represent the nucleic acid makeup of a single person. Reference genomes may be used to for mapping of sequencing reads from a sample to chromosomal positions. For example, a reference genome used for human subjects as well as many other organisms is found at the National Center for Biotechnology Information at ncbi.nlm.nih.gov.

[0051] The term "read segment" or "read" refers to any nucleotide sequences including sequence reads obtained from an individual and/or nucleotide sequences derived from the initial sequence read from a sample obtained from an individual.

[0052] The terms "sample," "patient sample," "biological sample," and the like, encompass a variety of sample types obtained from a patient, individual, or subject and can be used in a diagnostic, prognostic and/or monitoring assay. The patient sample may be obtained

from a healthy subject, a diseased patient, or a patient with lung cancer. In certain embodiments, a sample that is "provided" can be obtained by the person (or machine) conducting the assay, or it can have been obtained by another, and transferred to the person (or machine) carrying out the assay. Moreover, a sample obtained from a patient can be divided and only a portion may be used for diagnosis. Further, the sample, or a portion thereof, can be stored under conditions to maintain sample for later analysis. The definition specifically encompasses blood and other liquid samples of biological origin (including, but not limited to, peripheral blood, serum, plasma, cord blood, amniotic fluid, cerebrospinal fluid, urine, saliva, stool and synovial fluid), solid tissue samples such as a biopsy specimen or tissue cultures or cells derived therefrom and the progeny thereof. In certain embodiment, a sample comprises cerebrospinal fluid. In a specific embodiment, a sample comprises a blood sample. In another embodiment, a sample comprises a plasma sample. In yet another embodiment, a serum sample is used. The definition of "sample" also includes samples that have been manipulated in any way after their procurement, such as by centrifugation, filtration, precipitation, dialysis, chromatography, treatment with reagents, washed, or enriched for certain cell populations. The terms further encompass a clinical sample, and also include cells in culture, cell supernatants, tissue samples, organs, and the like. Samples may also comprise fresh-frozen and/or formalin-fixed, paraffin-embedded tissue blocks, such as blocks prepared from clinical or pathological biopsies, prepared for pathological analysis or study by immunohistochemistry.

[0053] The term "sequence reads" refers to nucleotide sequences read from a sample obtained from an individual. Sequence reads can be obtained through various methods known in the art.

[0054] As defined herein, a "therapeutically effective" amount of a compound or agent (i.e., an effective dosage) means an amount sufficient to produce a therapeutically (e.g., clinically) desirable result. The compositions can be administered from one or more times per day to one or more times per week, including once every other day. The skilled artisan will appreciate that certain factors can influence the dosage and timing required to effectively treat a subject, including but not limited to the severity of the disease or disorder, previous treatments, the general health and/or age of the subject, and other diseases present. Moreover, treatment of a

subject with a therapeutically effective amount of the compounds of the invention can include a single treatment or a series of treatments.

[0055] As used herein, the terms "treat," treating," "treatment," and the like refer to reducing or ameliorating a disorder and/or symptoms associated therewith. It will be appreciated that, although not precluded, treating a disorder or condition does not require that the disorder, condition or symptoms associated therewith be completely eliminated.

[0056] Genes: All genes, gene names, and gene products disclosed herein are intended to correspond to homologs from any species for which the compositions and methods disclosed herein are applicable. It is understood that when a gene or gene product from a particular species is disclosed, this disclosure is intended to be exemplary only, and is not to be interpreted as a limitation unless the context in which it appears clearly indicates. Thus, for example, for the genes or gene products disclosed herein, are intended to encompass homologous and/or orthologous genes and gene products from other species.

[0057] Ranges: throughout this disclosure, various aspects of the invention can be presented in a range format. It should be understood that the description in range format is merely for convenience and brevity and should not be construed as an inflexible limitation on the scope of the invention. Accordingly, the description of a range should be considered to have specifically disclosed all the possible subranges as well as individual numerical values within that range. For example, description of a range such as from 1 to 6 should be considered to have specifically disclosed subranges such as from 1 to 3, from 1 to 4, from 1 to 5, from 2 to 4, from 2 to 6, from 3 to 6 etc., as well as individual numbers within that range, for example, 1, 2, 2.7, 3, 4, 5, 5.3, and 6. This applies regardless of the breadth of the range.

[0058] Any compositions or methods provided herein can be combined with one or more of any of the other compositions and methods provided herein.

BRIEF DESCRIPTION OF THE DRAWINGS

[0059] The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

[0060] **FIGS. 1A-1C** demonstrate the enrichment of motifs at the ends of cfDNA-fragments and at recurrent cfDNA-fragment ends. The 3 bp motifs are located around the fragment, with one base outside the fragment and the first two bases of the fragment. The vertical line indicates the start of the fragment. **FIG. 1A:** Frequency of 3 bp DNA-motifs at the ends of DNA fragments after shearing by sonication, at the ends of cfDNA fragments, and at the ends of cfDNA with "preferred" ends that are observed at least in 5% of recurrent sequences. The relative frequencies are normalized by the observed occurrence of 3 bp motifs in the human genome. **FIG. 1B:** Increased preference of motifs in recurrent cfDNA-fragment end-positions, as measured by an increased ratio of the number of cfDNA-fragments starting or stopping at a certain position, over the total amount of cfDNA fragments overlapping a window of +/- 50 bases around that position. These included enrichment of T|CC and A|CC as well as T|CG and A|CG at these recurrent cfDNA end-positions. **FIG 1C:** X-ray crystal structure of PDB entry 7COW[44]. TCG motifs colored red, nucleosome protein shown in grey surface with the histone H1 linker at the top. Bases within 5 angstroms of H1 linker shown as spheres.

[0061] **FIGS. 2A-2E** demonstrate that DNA methylation is a determinant of cfDNA-fragmentation. a. The frequency of observed CpG's at different positions in cfDNA fragments, counted from the cfDNA break, differs between methylated and unmethylated CpG sites. While unmethylated CpG's show a more equal distribution over the cfDNA fragments, methylated CpG's show enrichment at the beginning of cfDNA-fragments. **FIG. 2B:** The preference of fragments to start at a CpG increases with higher levels of methylation of that CpG. The preference for cfDNA-fragments to start at a CpG is measured by the ratio of cfDNA-fragments starting with a CpG, over the total amount of cfDNA fragments overlapping a window of +/- 50 bases around the first base that CpG. **FIG. 2C:** The opposite relationship is seen when a CpG is preceded by a cytosine: there is a preference for cfDNA fragments to start with CCG, when the CpG in this motif is not methylated. **FIG. 2D:** CpG's on chromosome X are known to be differently methylated in male and female individuals. Due to X-inactivation by methylation of CpG-islands, these CpG's show a difference in the preference of cfDNA-fragment end-positions. With increased methylation, more fragments start with a CpG, while less fragments start with a

CCG. **FIG. 2E:** On autosomes, we do not observe a difference in methylation between male and female individuals.

[0062] **FIGS. 3A-3G** demonstrate the effect of CpG methylation and gene expression on coverage and size of cfDNA fragments. **FIG 3A:** Sequence coverage in regions of CpG-islands, ordered by average methylation of the CpG-island. **FIG 3B:** Sequence coverage in regions of transcription start sites, ordered by the average expression of genes in myeloid cell lines. **FIG. 3C:** Average cfDNA fragment sizes in regions of CpG-islands, ordered by average methylation of the CpG-island. **FIG. 3D:** Average cfDNA fragment sizes in regions of transcription start sites, ordered by the average expression of genes in myeloid cell lines. **FIG. 3E:** A correlation matrix (right) of significant gene-set enrichment analyses of gene-expression, DNA-methylation at the CpG-islands, cfDNA fragment coverage over the transcription start sites, and cfDNA coverage over the CpG-islands, all associated with the same genes. Representative gene set analyses (left) of KEGG neuronal receptor-ligand interactions, a gene set with reduced expression and increased methylation in WBC, showing increased cfDNA coverage at TSS and CpG sites. **FIG. 3F:** cfDNA fragment size distributions (top) and cumulative representations (bottom) for fragments with a mutation (tumor-derived) compared to wild-type fragments (mainly white blood cells), fragments from regions of high expression transcription start sites compared to fragments from regions of low expression transcription start sites, and fragments from regions of methylated CpG-islands compared to fragments from regions of unmethylated CpG-islands. **FIG.3G:** Data from human isogenic xenografts (IDH1 R132H mutant compared to IDH1 wild-type human glioblastoma cell lines) showing increased coverage of human cfDNA in areas of increased methylation and decreased coverage in areas of increased gene expression over the top n regions of differential methylation or gene expression

[0063] **FIGS. 4A, 4B** show a comparison of cfDNA fragment end motifs in regions of differential methylation in individuals with and without pancreatic cancer. **FIG. 4A:** Aggregated ratios of cfDNA fragments starting or ending at specific motifs containing CpG's which showed differential methylation between cfDNA of healthy individuals and pancreatic cancer tissues. The largest increase in signal was found in pancreatic cancer patients (n=34), while cfDNA from patients with other cancers (colorectal n=27, ovarian n=28, lung n=39, and breast cancer n=54)

showed intermediate signals between the signal found in pancreatic cancer patients and individuals without cancer (n=244). **FIG. 4B:** The predictive value of this signal for detecting pancreatic cancer is shown as receiver operating characteristic curve in comparison with the DELFI, while combining the DELFI approach with the methylation-based signal in an ensemble-model showed the best prediction.

[0064] **FIG. 5** demonstrates the relative frequencies of nucleotides over 144bp cfDNA fragments. The bar graphs show the relative frequency of nucleotides in all 144bp cfDNA fragments of 543 healthy individuals. The observed frequency shows a 10.4 base periodicity in the occurrence of nucleotides with A's and T's (more flexible) alternating with C's and G's (less flexible).

[0065] **FIG. 6** demonstrates that the fraction of cfDNA fragments starting or ending at specific CpG-containing motifs is dependent on the CpG methylation status. Increasing methylation of CpGs results in more fragments starting at that position while an opposite relationship is observed in fragments starting with CCG. This effect is independent of the genomic context in which the CpG occurs, including in CpG-islands, -shores, -shelves or in open sea.

[0066] **FIGS. 7A, 7B** show the amount of cfDNA fragments starting at motifs containing a CpG. The top two histograms (light blue and dark blue) show the normalized amount of fragments starting at each position around a motif, obtained from aggregating the low coverage WGS of cfDNA from 543 individuals without cancer. In the light blue histogram, the motif contains an unmethylated CpG (beta-value < 0.3). In the dark blue histogram, the motif contains a methylated CpG (beta-value > 0.7). In the third histogram (brown), the difference between the upper two histograms is shown, revealing an increase in fragments starting at a CpG when that CpG is methylated. When an extra cytosine is positioned in front of the CpG, we also observe a major decrease in the amount of cfDNA fragments starting at that cytosine, when the following CpG is methylated. **FIG. 7A** shows the effect of methylation on cfDNA fragments starting around 3bp motifs, while **FIG. 7B** shows the effect of methylation on fragments starting around 4bp motifs.

[0067] **FIG. 8** demonstrates that the cfDNA fragment ends are affected by differences in naturally occurring methylation on the X chromosome between male and female individuals due to X-inactivation. Similar to FIG. 2D and FIG. 2E which is focused on CpG-islands on the X chromosome and the autosomes, fragments starting at CpG-shores in the X chromosome are different between males and females due to increased methylation at these locations on the second chromosome X. In CpG-shelves and open sea CpGs on the X chromosome, an opposite pattern is observed. In these areas, female individuals show less fragments starting with CpGs and more fragments starting with CCGs, indicating less methylation of these CpGs in female X chromosomes. When a CpG is preceded by a cytosine, the opposite relationship is observed. As control, it was shown that on the autosomes the amount of cfDNA fragments starting at CpGs or CpGs preceded by a cytosine is not different between males and females.

[0068] **FIGS. 9A** and **9B** demonstrate the relationship between CpG-island methylation, gene expression, and cfDNA coverage. **FIG. 9A:** The dot-plot shows that decreased methylation of CpG-islands (lower beta-values) is associated with lower cfDNA coverage at CpG islands while increased methylation of CpG-islands (higher beta-values) is associated with a higher cfDNA coverage at CpG islands. **FIG. 9B** Gene expression shows an inverse relationship with cfDNA coverage around the transcription start sites where increased gene expression is related to decreased cfDNA coverage, while decreased gene expression shows increased cfDNA coverage.

[0069] **FIGS. 10A-10E** demonstrate the effect of methylation and gene expression on cfDNA coverage and fragment-size at large regions around CpG-islands and transcription start sites. **FIG. 10A:** Cumulative coverage and (**FIG. 10C**) average cfDNA-fragment size around the 1000 most methylated and 1000 least methylated CpG-islands (centered around the middle of the CpG-island), shows decreased coverage and smaller fragments at less methylated CpG-islands. **FIG. 10B:** Cumulative coverage and (**FIG. 10D**) average cfDNA fragment size around the1000 most expressed genes and 1000 least expressed genes (centered around the transcription start site), shows decreased coverage and smaller fragment sizes around more expressed genes. The effect is most pronounced directly around the transcription start site. **FIG. 10E:** Average methylation around the 1000 most methylated and 1000 least methylated CpG-islands (centered around the middle of the CpG-island).

[0070] **FIGS. 11A, 11B** demonstrate the effect of gene expression and CpG-island methylation on cfDNA coverage and on the fraction of cfDNA fragments starting or ending at CpGs. **FIG. 11A.** Cumulative cfDNA coverage around the transcription start sites of the top 1000 most expressed and top 1000 least expressed genes of all genes, and in subgroups of only among genes with unmethylated CpG-islands (beta-value < 0.3) or methylated CpG-islands (beta-value > 0.7). **FIG. 11B.** When genes are ordered by the amount of methylation of their associated CpG-islands (left side), the fraction of cfDNA fragments starting with CpGs in these genes show a similar increase. Ordering genes by expression (right side), shows a much more limited effect on the fraction of cfDNA fragments starting with CpGs in these genes. It was concluded that methylation has an independent effect in comparison to expression on the amount of fragments starting at CpGs.

[0071] **FIGS. 12A, 12B** show the number of expressed genes and sites of methylation in white blood cells and cancer tissues. **FIG 12A:** Healthy white blood-cells show statistically significant fewer expressed genes than cancer tissues. **FIG. 12B:** White blood-cells show statistically significant more methylated CpG-islands than cancer tissues. BRCA: breast cancer, COAD: colon adenocarcinoma, LIHC: liver hepatocellular carcinoma, LUAD: lung adenocarcinoma, LUSC: lung squamous cell carcinoma, OV: ovarian carcinoma, PAAD: pancreatic adenocarcinoma.

[0072] **FIG. 13** (includes **FIGS. 13A and 13B**): Multivariate analysis showing the impact of CpG-island methylation, RNA expression and nucleosome positioning on cfDNA coverage and size. **FIG. 13A**: Forest plot displaying regression coefficients of a model for coverage with nucleosome positioning score (WPS), RNA expression (RNA), and an indicator for methylation status (Meth = 1 if beta $\geq$ 0.5 and 0 otherwise) as independent predictors. Both RNA and WPS were scaled to have unit standard deviation. To model the non-linear relationship between coverage and WPS, quadratic (WPS^2) and cubic terms (WPS^3) were included. Additionally, we allowed the relationship between coverage and WPS to depend on the level of RNA expression and methylation status through the addition of interaction terms. An analysis of variance (ANOVA) comparing the model with only WPS and RNA terms (WPS + WPS^2 + WPS^3 + RNA + WPS x RNA + WPS^2 x RNA + WPS^3 x RNA) to the full model

shown in panel (a) was statistically significant ($F_{5,18377} = 22.0$, $p < 0.0001$), indicating that methylation was helpful for explaining variation in coverage beyond the effects of nucleosome positioning and RNA expression alone. Similarly, an ANOVA comparing the model with only WPS and Meth terms (WPS + WPS^2 + WPS^3 + Meth + WPS x Meth + WPS^2 x Meth + WPS^3 x Meth) to the full model demonstrated that RNA expression was statistically significant and helpful for explaining variation in coverage beyond the effects of WPS and methylation alone ($F_{5,18377} = 1271.2$, $p < 0.0001$). Corresponding ANOVAs with WPS as linear or quadratic models were also significant across analyses ($p<0.0001$). **FIG. 13B**: Forest plot displaying the results of the model with cfDNA fragment size (bp) as the dependent variable and WPS, RNA, and Meth as independent variables. Again, quadratic and cubic terms for WPS allowed for non-linearity between fragment size and WPS. ANOVA analyses comparing the full model (panel b.) to the model without methylation and the model without RNA expression were each statistically significant (Meth ANOVA: $F_{5,18539}$, $p = 1.3e\text{-}10$; RNA ANOVA: $F_{5,18539}$, $p < 0.0001$).

DETAILED DESCRIPTION

[0073] The disclosure herein, provides a connection between methylation, expression and cfDNA fragmentation in healthy subjects and subjects with cancer. Further provided herein, is an analysis of features related to motifs associated with nucleosome positioning and breakpoints of cfDNA fragments in both healthy individuals and patients with cancer. The disclosure also provides how epigenetic marks give rise to specific patterns of fragmentation of cfDNA and how these are related to both methylation and gene expression. In the examples section which follow, it is demonstrated that differentially methylated CpG's in specific sequence contexts can be used to identify differences in cfDNA fragmentation between healthy individuals and cancer patients.

[0074] ***CpG Methylation***

[0075] Epigenetics refers to heritable changes in gene expression that are not due to changes in DNA. The best defined epigenetic change is DNA methylation of cytosines, by DNA methyltransferase enzymes. Cytosines associated with guanines are called CpG dinucleotides, and these are generally found in CpG-rich regions called CpG islands. CpG islands are defined as sequence ranges with at least 200 bp, a GC percentage greater than 50%, and an observed-to-

expected (Obs/Exp) CpG ratio greater than 60%. The expected number of CpG dimers in a window is calculated as the number of 'C's in the window multiplied by the number of 'G's in the window, divided by the window length. CpG islands are defined as regions of greater than 500 bp that have guanine cytosine content of greater than 55%. Up to 60% of CpG islands are in the 5′ regulatory (promoter) regions of genes. However, CpG islands that are not in promoter regions can also be found within coding regions and noncoding regions of genes, which may be targets for *de novo* methylation in cancer and aging. DNA methylation affects a number of different cellular processes including apoptosis, cell cycle, DNA damage repair, growth factor response, signal transduction, and tumor architecture, all of which can contribute to the initiation and progression of cancer.

[0076] Methylated cytosines can be in CpG islands, shores, shelves, open sea, and sites surrounding transcription sites (-200 to -1500 bp, 5′ untranslated region (UTR), and exons 1]) for coding genes as well as gene bodies and 3′UTR and other/open sea regions derived from genome-wide association studies. Shores are considered regions 0–2 kb from CpG islands, shelves are regions 2–4 kb from CpG islands, and other/open sea regions are isolated CpG sites in the genome that do not have a specific designation.

[0077] Accordingly, in certain embodiments, genomic sequences are assayed to identify genome-wide CpG methylation and frequency of cfDNA breakpoints at a plurality of positions in the genomic sequences to determine circulating cell-free DNA (cfDNA) fragmentation, wherein, recurrent cfDNA fragment end enrichment at CpG sites correlate with higher genome-wide methylation levels and smaller cfDNA fragments diagnostic of cancer.

[0078] In certain embodiments, the method further comprises mapping cfDNA fragments to the genome and comparing the cfDNA fragment end sequences to methylated and unmethylated CpG sites of cfDNA from healthy subjects. The methylated CpGs are enriched at the ends of A/T|CG cfDNA fragment sequence and quantitative assessment of enrichment of cfDNA fragment ends at CpGs comprise calculating for each CpG a fraction of cfDNA fragments starting or ending at CpG dinucleotide positions over number of cfDNA fragments with start or end positions within 50 bp around each CpG. In certain embodiments, cfDNA fragments obtained from an X chromosome among healthy subjects comprise increased cfDNA

fragments ending with CG and cfDNA fragment ends with CCG are decreased at locations of X chromosome CpG islands in female subjects as compared to male subjects. In certain embodiments, cfDNA sequence coverage of enriched cfDNA fragment end sequences comprise increased methylation across regions of methylated CpG islands as compared to reduced or lower occurrence cfDNA fragment end sequences. In certain embodiments, gene expression at transcription start sites (TSS) is inversely related to cfDNA coverage at TSS.

[0079] ***cfDNA Fragmentation Profiles***

[0080] A cfDNA fragmentation profile can include one or more cfDNA fragmentation patterns. A cfDNA fragmentation pattern can include any appropriate cfDNA fragmentation pattern. . Examples of cfDNA fragmentation patterns include, without limitation, median fragment size, fragment size distribution, ratio of small cfDNA fragments to large cfDNA fragments, and the coverage of cfDNA fragments. In some embodiments, a cfDNA fragmentation pattern includes two or more (e.g., two, three, or four) of median fragment size, fragment size distribution, ratio of small cfDNA fragments to large cfDNA fragments, and the coverage of cfDNA fragments. In some embodiments, cfDNA fragmentation profile can be a genome-wide cfDNA profile (e.g., a genome-wide cfDNA profile in windows across the genome). In some embodiments, cfDNA fragmentation profile can be a targeted region profile. A targeted region can be any appropriate portion of the genome (e.g., a chromosomal region). Examples of chromosomal regions for which a cfDNA fragmentation profile can be determined as described herein include, without limitation, a portion of a chromosome (e.g., a portion of 2q, 4p, 5p, 6q, 7p, 8q, 9q, 10q, 11q, 12q, and/or 14q) and a chromosomal arm (e.g., a chromosomal arm of 8q,13q, 11q, and/or 3p). In some embodiments, a cfDNA fragmentation profile can include two or more targeted region profiles.

[0081] In some embodiments, a cfDNA fragmentation profile can be used to identify changes (e.g., alterations) in cfDNA fragment lengths. An alteration can be a genome-wide alteration or an alteration in one or more targeted regions/loci. A target region can be any region containing one or more cancer-specific alterations. In some embodiments, a cfDNA fragmentation profile can be used to identify (e.g., simultaneously identify) from about 10 alterations to about 500 alterations (e.g., from about 25 to about 500, from about 50 to about 500,

from about 100 to about 500, from about 200 to about 500, from about 300 to about 500, from about 10 to about 400, from about 10 to about 300, from about 10 to about 200, from about 10 to about 100, from about 10 to about 50, from about 20 to about 400, from about 30 to about 300, from about 40 to about 200, from about 50 to about 100, from about 20 to about 100, from about 25 to about 75, from about 50 to about 250, or from about 100 to about 200, alterations).

[0082] A cfDNA fragmentation profile can be obtained using any appropriate method. In some embodiments, cfDNA from a mammal (e.g., a mammal having, or suspected of having, cancer) can be processed into sequencing libraries which can be subjected to whole genome sequencing (e.g., low-coverage whole genome sequencing), mapped to the genome, and analyzed to determine cfDNA fragment lengths. Mapped sequences can be analyzed in non-overlapping windows covering the genome. Windows can be any appropriate size. For example, windows can be from thousands to millions of bases in length. As one non-limiting example, a window can be about 5 megabases (Mb) long. Any appropriate number of windows can be mapped. For example, tens to thousands of windows can be mapped in the genome. For example, hundreds to thousands of windows can be mapped in the genome. A cfDNA fragmentation profile can be determined within each window.

[0083] In some embodiments, methods and materials described herein also can include machine learning. For example, machine learning can be used for identifying mutation frequencies, altered fragmentation profile (e.g., using coverage of cfDNA fragments, fragment size of cfDNA fragments, coverage of chromosomes, and mtDNA).

[0084] In some embodiments, determining a cfDNA fragmentation profile in a mammal can be used for identifying a mammal as having cancer. For example, cfDNA fragments obtained from a mammal (e.g., from a sample obtained from a mammal) can be subjected to low coverage whole-genome sequencing, and the sequenced fragments can be mapped to the genome and assessed to determine a cfDNA fragmentation profile. As described herein, a cfDNA fragmentation profile of a mammal having cancer is more heterogeneous (e.g., in fragment lengths) than a cfDNA fragmentation profile of a healthy mammal (e.g., a mammal not having cancer). As such, also provided are methods and materials for assessing, monitoring, and/or treating mammals (e.g., humans) having, or suspected of having, cancer. In some embodiments,

methods and materials are provided for identifying a mammal as having cancer. For example, a sample (e.g., a blood sample) obtained from a mammal can be assessed to determine the presence and, optionally, the tissue of origin of the cancer in the mammal based, at least in part, on the cfDNA fragmentation profile of the mammal. In some embodiments, methods and materials are provided for monitoring a mammal as having cancer. For example, a sample (e.g., a blood sample) obtained from a mammal can be assessed to determine the presence of the cancer in the mammal based, at least in part, on the cfDNA fragmentation profile of the mammal. In some embodiments, methods and materials are provided for identifying a mammal as having cancer and administering one or more cancer treatments to the mammal to treat the mammal. For example, a sample (e.g., a blood sample) obtained from a mammal can be assessed to determine if the mammal has cancer based, at least in part, on the cfDNA fragmentation profile of the mammal, and one or more cancer treatments can be administered to the mammal.

[0085] In some embodiments, a cfDNA fragmentation profile can be used to detect tumor-derived DNA. For example, a cfDNA fragmentation profile can be used to detect tumor-derived DNA by comparing a cfDNA fragmentation profile of a mammal having, or suspected of having, cancer to a reference cfDNA fragmentation profile (e.g., a cfDNA fragmentation profile of a healthy mammal and/or a nucleosomal DNA fragmentation profile of healthy cells from the mammal having, or suspected of having, cancer). In some embodiments, a reference cfDNA fragmentation profile is a previously generated profile from a healthy mammal. For example, methods provided herein can be used to determine a reference cfDNA fragmentation profile in a healthy mammal, and that reference cfDNA fragmentation profile can be stored (e.g., in a computer or other electronic storage medium) for future comparison to a test cfDNA fragmentation profile in mammal having, or suspected of having, cancer. In some embodiments, a reference cfDNA fragmentation profile (e.g., a stored cfDNA fragmentation profile) of a healthy mammal is determined over the whole genome. In some embodiments, a reference cfDNA fragmentation profile (e.g., a stored cfDNA fragmentation profile) of a healthy mammal is determined over a subgenomic interval.

[0086] In some embodiments, a cfDNA fragmentation profile can be used to identify a mammal (e.g., a human) as having cancer (e.g., a liver cancer, a colorectal cancer, a lung cancer,

a breast cancer, a gastric cancer, a pancreatic cancer, a bile duct cancer, and/or an ovarian cancer).

[0087] A cfDNA fragmentation profile can include a cfDNA fragment size pattern. cfDNA fragments can be any appropriate size. For example, cfDNA fragment can be from about 50 base pairs (bp) to about 400 bp in length.

[0088] A cfDNA fragmentation profile can include a cfDNA fragment size distribution. As described herein, a mammal having cancer can have a cfDNA size distribution that is more variable than a cfDNA fragment size distribution in a healthy mammal. In some embodiments, a size distribution can be within a targeted region. A healthy mammal (e.g., a mammal not having cancer) can have a targeted region cfDNA fragment size distribution of about 1 or less than about 1. In some embodiments, a mammal having cancer can have a targeted region cfDNA fragment size distribution that is longer (e.g., 10, 15, 20, 25, 30, 35, 40, 45, 50 or more bp longer, or any number of base pairs between these numbers) than a targeted region cfDNA fragment size distribution in a healthy mammal. In some embodiments, a mammal having cancer can have a targeted region cfDNA fragment size distribution that is shorter (e.g., 10, 15, 20, 25, 30, 35, 40, 45, 50 or more bp shorter, or any number of base pairs between these numbers) than a targeted region cfDNA fragment size distribution in a healthy mammal. In some embodiments, a size distribution can be a genome-wide size distribution. A healthy mammal (e.g., a mammal not having cancer) can have very similar distributions of short and long cfDNA fragments genome wide. In some embodiments, a mammal having cancer can have, genome-wide, one or more alterations (e.g., increases and decreases) in cfDNA fragment sizes. The one or more alterations can be any appropriate chromosomal region of the genome. For example, an alteration can be in a portion of a chromosome. Examples of portions of chromosomes that can contain one or more alterations in cfDNA fragment sizes include, without limitation, portions of 2q, 4p, 5p, 6q, 7p, 8q, 9q, 10q, 11q, 12q, and 14q. For example, an alteration can be across a chromosome arm (e.g., an entire chromosome arm).

[0089] A cfDNA fragmentation profile can include a ratio of small cfDNA fragments to large cfDNA fragments and a correlation of fragment ratios to reference fragment ratios. As used herein, with respect to ratios of small cfDNA fragments to large cfDNA fragments, a small

cfDNA fragment can be from about 100 bp in length to about 150 bp in length. As used herein, with respect to ratios of small cfDNA fragments to large cfDNA fragments, a large cfDNA fragment can be from about 151 bp in length to 220 bp in length. A mammal having cancer can have a correlation of fragment ratios (e.g., a correlation of cfDNA fragment ratios to reference DNA fragment ratios such as DNA fragment ratios from one or more healthy mammals) that is lower (e.g., 2-fold lower, 3-fold lower, 4-fold lower, 5-fold lower, 6-fold lower, 7-fold lower, 8-fold lower, 9-fold lower, 10-fold lower, or more) than in a healthy mammal. A healthy mammal (e.g., a mammal not having cancer) can have a correlation of fragment ratios (e.g., a correlation of cfDNA fragment ratios to reference DNA fragment ratios such as DNA fragment ratios from one or more healthy mammals) of about 1 (e.g., about 0.96). In some embodiments, a mammal having cancer can have a correlation of fragment ratios (e.g., a correlation of cfDNA fragment ratios to reference DNA fragment ratios such as DNA fragment ratios from one or more healthy mammals) that is, on average, lower than a correlation of fragment ratios (e.g., a correlation of cfDNA fragment ratios to reference DNA fragment ratios such as DNA fragment ratios from one or more healthy mammals) in a healthy mammal.

[0090] A cfDNA fragmentation profile can include coverage of all fragments. Coverage of all fragments can include windows (e.g., non-overlapping windows) of coverage. In some embodiments, coverage of all fragments can include windows of small fragments (e.g., fragments from about 100 bp to about 150 bp in length). In some embodiments, coverage of all fragments can include windows of large fragments (e.g., fragments from about 151 bp to about 220 bp in length).

[0091] In certain embodiments, a cfDNA fragmentation profile can be used to identify the molecular origins of cfDNA in patients and identify genomic and chromatin features associated with fragmentation changes.

[0092] In some embodiments, a cfDNA fragmentation profile can be used to identify the tissue of origin of a cancer (e.g., a liver cancer, a colorectal cancer, a lung cancer, a breast cancer, a gastric cancer, a pancreatic cancer, a bile duct cancer, or an ovarian cancer). For example, a cfDNA fragmentation profile can be used to identify a localized cancer. When a cfDNA fragmentation profile includes a targeted region profile, one or more alterations

described herein can be used to identify the tissue of origin of a cancer. In some embodiments, one or more alterations in chromosomal regions can be used to identify the tissue of origin of a cancer.

[0093] A cfDNA fragmentation profile can be obtained using any appropriate method. In some embodiments, cfDNA from a mammal (e.g., a mammal having, or suspected of having, cancer) can be processed into sequencing libraries which can be subjected to whole genome sequencing (e.g., low-coverage whole genome sequencing), mapped to the genome, and analyzed to determine cfDNA fragment lengths. Mapped sequences can be analyzed in non-overlapping windows covering the genome. Windows can be any appropriate size. For example, windows can be from thousands to millions of bases in length. As one non-limiting example, a window can be about 5 megabases (Mb) long. Any appropriate number of windows can be mapped. For example, tens to thousands of windows can be mapped in the genome. For example, hundreds to thousands of windows can be mapped in the genome. A cfDNA fragmentation profile can be determined within each window.

[0094] In some embodiments, methods and materials described herein also can include machine learning. For example, machine learning can be used for identifying an altered fragmentation profile (e.g., using coverage of cfDNA fragments, fragment size of cfDNA fragments, coverage of chromosomes, and mtDNA). Various machine learning algorithms can be used to analyze the fragmentation profiles. For example, to distinguish healthy from cancer patients using fragmentation profiles, a stochastic gradient boosting model is used (gbm; see, e.g., Friedman *et al.*, 2001 *Ann Stat* 29:1189-1232; and Friedman *et al.*, 2002 *Comput Stat Data An* 38:367-378). GC-corrected total and short fragment coverage for all 504 bins can be centered and scaled for each sample to have mean 0 and unit standard deviation. Additional features included Z-scores for each of the 39 autosomal arms and mitochondrial representation (log10-transformed proportion of reads mapped to the mitochondria). To estimate the prediction error of this approach, 10-fold cross-validation can be used as described elsewhere (see, e.g., Efron et al., 1997 J Am Stat Assoc 92, 548-560). Feature selection, performed only on the training data in each cross-validation run, removed bins that were highly correlated (correlation > 0.9) or had near zero variance. Stochastic gradient boosted machine learning can be implemented using the

R package gbm package. To average over the prediction error from the randomization of patients to folds, the 10-fold cross validation procedure can be repeated.

[0095] In some embodiments, a machine learning model is a neural network (NN). In certain embodiments, the neural network is a convolutional neural network, a recurrent neural network, or a deep learning neural network. In some embodiments, the machine learning model is a random forest, logistic regression, or an unsupervised clustering model. In other embodiments of neural network models, the model can also be a deep neural network (DNN) with multiple locally and fully connected hidden layers, or a high-order neural network (HONN). For DNN, a Restricted Boltzmann Machine (RBM) can be used to pre-train the neural nodes of input and connecting layers. For HONN, a mean-covariance RBM can be used to pre-train the neural nodes of input and connecting layers.

[0096] In certain embodiments, a computer system for obtaining access to database files and executing one or more software programs may include a server, a data storage device, a network, and a user interface device. The server may also be a hypervisor-based system executing one or more guest partitions hosting operating systems with modules having server configuration information. In a further embodiment, the system may include a storage controller, or a storage server configured to manage data communications between the data storage device and the server or other components in communication with the network. In an alternative embodiment, the storage controller may be coupled to the network.

[0097] In certain embodiments, the user interface device is referred to broadly and is intended to encompass a suitable processor-based device such as a desktop computer, a laptop computer, a personal digital assistant (PDA) or tablet computer, a smartphone or other mobile communication device having access to the network. In a further embodiment, the user interface device may access the Internet or other wide area or local area network to access a web application or web service hosted by the server and may provide a user interface for enabling a user to enter or receive information. The network may facilitate communications of data between the server and the user interface device. The network may include any type of communications network including, but not limited to, a direct PC-to-PC connection, a local area network (LAN), a wide area network (WAN), a modem-to-modem connection, the Internet, a combination of the

above, or any other communications network now known or later developed within the networking arts which permits two or more computers to communicate.

[0098] In certain embodiments, a computer system comprises a central processing unit ("CPU") coupled to the system bus. The CPU may be a general purpose CPU or microprocessor, graphics processing unit ("GPU"), and/or microcontroller. The CPU may execute the various logical instructions according to the present embodiments. The computer system may also include random access memory (RAM), which may be synchronous RAM (SRAM), dynamic RAM (DRAM), synchronous dynamic RAM (SDRAM), or the like. The computer system may utilize RAM to store the various data structures used by a software application. The computer system may also include read only memory (ROM) which may be PROM, EPROM, EEPROM, optical storage, or the like. The ROM may store configuration information for booting the computer system. The RAM and the ROM hold user and system data, and both the RAM and the ROM may be randomly accessed.

[0099] The computer system may also include an I/O adapter, a communications adapter, a user interface adapter, and a display adapter. The I/O adapter and/or the user interface adapter may, in certain embodiments, enable a user to interact with the computer system. In a further embodiment, the display adapter may display a graphical user interface (GUI) associated with a software or web-based application on a display device, such as a monitor or touch screen.

[00100] The I/O adapter may couple one or more storage devices, such as one or more of a hard drive, a solid state storage device, a flash drive, a compact disc (CD) drive, a floppy disk drive, and a tape drive, to the computer system. The data storage may be a separate server coupled to the computer system through a network connection to the I/O adapter. The communications adapter may be adapted to couple the computer system to the network, which may be one or more of a LAN, WAN, and/or the Internet. The user interface adapter couples user input devices, such as a keyboard, a pointing device, and/or a touch screen to the computer system. The display adapter may be driven by the CPU to control the display on the display device.

[00101] The computer system is provided as an example of one type of computing device that may be adapted to perform the functions of the server and/or the user interface device. For

example, any suitable processor-based device may be utilized including, without limitation, personal data assistants (PDAs), tablet computers, smartphones, computer game consoles, and multi-processor servers to implement various embodiments and/or steps the cancer detection models disclosed herein. Moreover, various embodiments of the cancer detection methods of the present disclosure may be implemented on application specific integrated circuits (ASIC), very large scale integrated (VLSI) circuits, or other circuitry. In fact, persons of ordinary skill in the art may utilize any number of suitable structures capable of executing logical operations according to the described embodiments. For example, the computer system may be virtualized for access by multiple users and/or applications.

[00102] Various methods, steps, calculations of parameters disclosed herein if implemented in firmware and/or software, the various functions described above may be stored as one or more instructions or code on a computer-readable medium. Examples include non-transitory computer-readable media encoded with a data structure and computer-readable media encoded with a computer program. Computer-readable media includes physical computer storage media. A storage medium may be any available medium that can be accessed by a computer. By way of example, and not limitation, such computer-readable media can comprise RAM, ROM, EEPROM, CD-ROM or other optical disk storage, magnetic disk storage or other magnetic storage devices, or any other medium, *e.g.*, cloud based, that can be used to store desired program code in the form of instructions or data structures and that can be accessed by a computer. Disk and disc include compact discs (CD), laser discs, optical discs, digital versatile discs (DVD), floppy disks and blu-ray discs. Generally, disks reproduce data magnetically, and discs reproduce data optically. Combinations of the above should also be included within the scope of computer-readable media.

[00103] In addition to storage on computer-readable medium, instructions and/or data may be provided as signals on transmission media included in a communication apparatus. For example, a communication apparatus may include a transceiver having signals indicative of instructions and data. The instructions and data are configured to cause one or more processors to implement the functions embodied herein.

[00104] **Methods of Treatment**

[00105] The methods embodied herein, include identifying a mammal as having cancer. The methods include, extracting cell-free DNA (cfDNA) from a subject's biological sample; generating genomic libraries from the extracted cfDNA; sequencing individual cfDNA molecules to obtain fragmentation profiles; analyze the features related to motifs associated with nucleosome positioning and breakpoints of cfDNA fragments in both healthy individuals and patients with cancer; analyze epigenetic marks give rise to specific patterns of fragmentation of cfDNA and how these are related to both methylation and gene expression. Using this information, differentially methylated CpG's in specific sequence contexts can be used to identify differences in cfDNA fragmentation between healthy individuals and cancer patients.

[00106] In another aspect, a method of diagnosing cancer and treating a subject, comprises obtaining a sample from a subject, assaying for changes in circulating cell-free DNA (cfDNA) fragment sizes as compared to normal and tumor-derived cfDNA controls, assaying for CpG methylation on cfDNA fragment ends and assessing fragment end representation at CG and CCG sites through low coverage whole genome cfDNA analyses; diagnosing the subject with cancer; and, treating the subject. In certain embodiments, an increase in cfDNA fragments ending in N|CCG as compared to a normal control is diagnostic of cancer. In certain embodiments, a decrease in cfDNA fragments ending with N|CCG is indicative of a subject without cancer. In certain embodiments, an increase of cfDNA fragments ending with CG as compared to normal controls is diagnostic of cancer. In certain embodiments, the cfDNA fragments are obtained from genomic regions with increased methylation as compared to normal genomic methylation controls. In certain embodiments, further comprises incorporating distribution of fragment end positions at CG and CCG sites in a gradient boosted tree machine learning model.

[00107] In another aspect, a method for diagnosing and treating a subject diagnosed with cancer, comprises assaying genomic sequences to identify genome-wide CpG methylation; analyzing frequency of cfDNA breakpoints at a plurality of positions in the genomic sequences and determining circulating cell-free DNA (cfDNA) fragmentation, wherein, recurrent cfDNA fragment end enrichment at CpG sites correlate with higher genome-wide methylation levels and smaller cfDNA fragments diagnostic of cancer; diagnosing the subject with cancer if hypomethylation and/or increased gene expression or a decrease in cfDNA fragment size is

detected; and, treating the subject. In certain embodiments, the genomic sequences are assayed by whole genome sequencing or obtaining whole genome sequences from a database and pooling cfDNA sequences. In certain embodiments, analysis of the frequency of cfDNA breakpoints comprises calculating a ratio of number of cfDNA fragments starting or ending at a particular position compared to the number of fragments with start or end positions within 50 bp surrounding that particular position. In certain embodiments, cfDNA fragments comprising similar end-position sequences comprise similar motifs. In certain embodiments, the motifs comprise a thymine or an adenine before a start of the cfDNA fragment sequence and two cytosines (A/T|CC) or a cytosine followed by a guanine (A/T|CG) as first two nucleotides of the cfDNA fragment sequence. In certain embodiments, the frequency of the motifs is increased in healthy subjects as compared to a subject having cancer. In certain embodiments, the frequency of A/T|CC motifs is greater than the frequency of A/T|CG motifs. In certain embodiments, the A/T|CG motifs are positioned close to a histone H1 linker or centered between 100 base pairs to 200 base pairs from the histone H1 linker. In certain embodiments, interior regions of cfDNA fragments are enriched with adenines and thymines. In certain embodiments, the method further comprises mapping cfDNA fragments to the genome and comparing the cfDNA fragment end sequences to methylated and unmethylated CpG sites of cfDNA from healthy subjects. In certain embodiments, methylated CpGs are enriched at the ends of A/T|CG cfDNA fragment sequence.

[00108] In certain embodiments, quantitative assessment of enrichment of cfDNA fragment ends at CpGs comprises calculating for each CpG a fraction of cfDNA fragments starting or ending at CpG dinucleotide positions over number of cfDNA fragments with start or end positions within 50 bp around each CpG. In certain embodiments, cfDNA fragments obtained from an X chromosome among healthy subjects comprise increased cfDNA fragments ending with CG and cfDNA fragment ends with CCG are decreased at locations of X chromosome CpG islands in female subjects as compared to male subjects. In certain embodiments, cfDNA sequence coverage of enriched cfDNA fragment end sequences comprise increased methylation across regions of methylated CpG islands as compared to reduced or lower occurrence cfDNA fragment end sequences. The mammal can have previously been

administered a cancer treatment to treat the cancer. The method also can include monitoring the mammal for the presence of cancer after administration of the cancer treatment.

[00109] In certain embodiments, a subject is diagnosed as having cancer, *e.g.* early stage cancer. In certain embodiments, the type and stage of cancer is identified and the subject is treated with one or more cancer therapies.

[00110] In some embodiments, methods and materials described herein for assessing, monitoring, and/or treating mammals (e.g., humans) having, or suspected of having, cancer. In some embodiments, methods and materials are provided for identifying a mammal as having cancer.

[00111] In some embodiments, methods and materials are provided for identifying a mammal as having cancer and administering one or more treatments to the mammal to treat the mammal. In some embodiments, during or after the course of a cancer treatment (e.g., any of the cancer treatments described herein), a mammal can undergo monitoring (or be selected for increased monitoring) and/or further diagnostic testing. In some embodiments, monitoring can include assessing mammals having, or suspected of having, cancer by, for example, assessing a sample (e.g., a blood sample) obtained from the mammal to determine features related to motifs associated with nucleosome positioning and breakpoints of cfDNA fragments in both healthy individuals and patients with cancer; analyze epigenetic marks give rise to specific patterns of fragmentation of cfDNA and how these are related to both methylation and gene expression. Using this information, differentially methylated CpG's in specific sequence contexts are used to identify differences in cfDNA fragmentation between healthy individuals and cancer patients to identify response to treatment and/or identify the mammal as having cancer (e.g., a residual cancer).

[00112] Any appropriate mammal can be assessed, monitored, and/or treated as described herein. A mammal can be a mammal having liver cancer. A mammal can be a mammal suspected of having liver cancer. Examples of mammals that can be assessed, monitored, and/or treated as described herein include, without limitation, humans, primates such as monkeys, dogs, cats, horses, cows, pigs, sheep, mice, and rats.

[00113] Any appropriate sample from a mammal can be assessed as described herein (e.g., assessed for a DNA fragmentation pattern). In some embodiments, a sample can include DNA (e.g., genomic DNA). In some embodiments, a sample can include cfDNA (e.g., circulating tumor DNA (ctDNA)). In some embodiments, a sample can be fluid sample (e.g., a liquid biopsy). Examples of samples that can contain DNA and/or polypeptides include, without limitation, blood (e.g., whole blood, serum, or plasma), amnion, tissue, urine, cerebrospinal fluid, saliva, sputum, broncho-alveolar lavage, bile, lymphatic fluid, cyst fluid, stool, ascites, pap smears, breast milk, and exhaled breath condensate.

[00114] A sample from a mammal to be assessed as described herein (ecan include any appropriate amount of cfDNA. In some embodiments, a sample can include a limited amount of DNA. For example, a cfDNA fragmentation profile can be obtained from a sample that includes less DNA than is typically required for other cfDNA analysis methods, such as those described in, for example, Phallen *et al.*, 2017 *Sci Transl Med* 9; Cohen *et al.*, 2018 *Science* 359:926; Newman *et al.*, 2014 *Nat Med* 20:548; and Newman *et al.*, 2016 *Nat Biotechnol* 34:547).

[00115] In some embodiments, a sample can be processed (e.g., to isolate and/or purify DNA and/or polypeptides from the sample). For example, DNA isolation and/or purification can include cell lysis (e.g., using detergents and/or surfactants), protein removal (e.g., using a protease), and/or RNA removal (e.g., using an RNase). As another example, polypeptide isolation and/or purification can include cell lysis (e.g., using detergents and/or surfactants), DNA removal (e.g., using a DNase), and/or RNA removal (e.g., using an RNase).

[00116] A cancer can be any stage cancer. In some embodiments, a cancer can be an early-stage cancer. In some embodiments, a cancer can be an asymptomatic cancer. In some embodiments, a cancer can be a residual disease and/or a recurrence (e.g., after surgical resection and/or after cancer therapy). A cancer can be any type of cancer. Examples of types of cancers that can be assessed, monitored, and/or treated as described herein include, without limitation, colorectal cancers, lung cancers, breast cancers, gastric cancers, pancreatic cancers, bile duct cancers, and ovarian cancers.

[00117] When treating a mammal having, or suspected of having, liver cancer as described herein, the mammal can be administered one or more cancer treatments. A cancer treatment can

be any appropriate cancer treatment. One or more cancer treatments described herein can be administered to a mammal at any appropriate frequency (e.g., once or multiple times over a period of time ranging from days to weeks). Examples of cancer treatments include, without limitation adjuvant chemotherapy, neoadjuvant chemotherapy, radiation therapy, hormone therapy, cytotoxic therapy, immunotherapy, adoptive T cell therapy (e.g., chimeric antigen receptors and/or T cells having wild-type or modified T cell receptors), targeted therapy such as administration of kinase inhibitors (e.g., kinase inhibitors that target a particular genetic lesion, such as a translocation or mutation), (e.g. a kinase inhibitor, an antibody, a bispecific antibody), signal transduction inhibitors, bispecific antibodies or antibody fragments (e.g., BiTEs), monoclonal antibodies, immune checkpoint inhibitors, surgery (e.g., surgical resection), or any combination of the above. In some embodiments, a cancer treatment can reduce the severity of the cancer, reduce a symptom of the cancer, and/or to reduce the number of cancer cells present within the mammal.

[00118] In some embodiments, a cancer treatment can include an immune checkpoint inhibitor. Non-limiting examples of immune checkpoint inhibitors include nivolumab (Opdivo), pembrolizumab (Keytruda), atezolizumab (tecentriq), avelumab (bavencio), durvalumab (imfinzi), ipilimumab (yervoy).

[00119] Cancer therapies in general also include a variety of combination therapies with both chemical and radiation based treatments. Combination chemotherapies include, for example, cisplatin (CDDP), carboplatin, procarbazine, mechlorethamine, cyclophosphamide, camptothecin, ifosfamide, melphalan, chlorambucil, busulfan, nitrosurea, dactinomycin, daunorubicin, doxorubicin, bleomycin, plicomycin, mitomycin, etoposide (VP16), tamoxifen, raloxifene, estrogen receptor binding agents, taxol, gemcitabien, navelbine, famesyl-protein transferase inhibitors, transplatinum, 5-fluorouracil, vincristine, vinblastine and methotrexate, Temazolomide (an aqueous form of DTIC), or any analog or derivative variant of the foregoing. The combination of chemotherapy with biological therapy is known as biochemotherapy. The chemotherapy may also be administered at low, continuous doses which is known as metronomic chemotherapy.

[00120] Yet further combination chemotherapies include, for example, alkylating agents such as thiotepa and cyclosphosphamide; alkyl sulfonates such as busulfan, improsulfan and piposulfan; aziridines such as benzodopa, carboquone, meturedopa, and uredopa; ethylenimines and methylamelamines including altretamine, triethylenemelamine, trietylenephosphoramide, triethiylenethiophosphoramide and trimethylolomelamine; acetogenins (especially bullatacin and bullatacinone); a camptothecin (including the synthetic analogue topotecan); bryostatin; callystatin; CC-1065 (including its adozelesin, carzelesin and bizelesin synthetic analogues); cryptophycins (particularly cryptophycin 1 and cryptophycin 8); dolastatin; duocarmycin (including the synthetic analogues, KW-2189 and CB1-TM1); eleutherobin; pancratistatin; a sarcodictyin; spongistatin; nitrogen mustards such as chlorambucil, chlornaphazine, cholophosphamide, estramustine, ifosfamide, mechlorethamine, mechlorethamine oxide hydrochloride, melphalan, novembichin, phenesterine, prednimustine, trofosfamide, uracil mustard; nitrosureas such as carmustine, chlorozotocin, fotemustine, lomustine, nimustine, and ranimnustine; antibiotics such as the enediyne antibiotics (e.g., calicheamicin, especially calicheamicin gammall and calicheamicin omegall; dynemicin, including dynemicin A; bisphosphonates, such as clodronate; an esperamicin; as well as neocarzinostatin chromophore and related chromoprotein enediyne antiobiotic chromophores, aclacinomysins, actinomycin, authrarnycin, azaserine, bleomycins, cactinomycin, carabicin, carminomycin, carzinophilin, chromomycinis, dactinomycin, daunorubicin, detorubicin, 6-diazo-5-oxo-L-norleucine, doxorubicin (including morpholino-doxorubicin, cyanomorpholino-doxorubicin, 2-pyrrolino-doxorubicin and deoxydoxorubicin), epirubicin, esorubicin, idarubicin, marcellomycin, mitomycins such as mitomycin C, mycophenolic acid, nogalarnycin, olivomycins, peplomycin, potfiromycin, puromycin, quelamycin, rodorubicin, streptonigrin, streptozocin, tubercidin, ubenimex, zinostatin, zorubicin; anti-metabolites such as methotrexate and 5-fluorouracil (5-FU); folic acid analogues such as denopterin, pteropterin, trimetrexate; purine analogs such as fludarabine, 6-mercaptopurine, thiamiprine, thioguanine; pyrimidine analogs such as ancitabine, azacitidine, 6-azauridine, carmofur, cytarabine, dideoxyuridine, doxifluridine, enocitabine, floxuridine; androgens such as calusterone, dromostanolone propionate, epitiostanol, mepitiostane, testolactone; anti-adrenals such as mitotane, trilostane; folic acid replenisher such

as frolinic acid; aceglatone; aldophosphamide glycoside; aminolevulinic acid; eniluracil; amsacrine; bestrabucil; bisantrene; edatraxate; defofamine; demecolcine; diaziquone; elformithine; elliptinium acetate; an epothilone; etoglucid; gallium nitrate; hydroxyurea; lentinan; lonidainine; maytansinoids such as maytansine and ansamitocins; mitoguazone; mitoxantrone; mopidanmol; nitraerine; pentostatin; phenamet; pirarubicin; losoxantrone; podophyllinic acid; 2-ethylhydrazide; procarbazine; PSK polysaccharide complex; razoxane; rhizoxin; sizofiran; spirogermanium; tenuazonic acid; triaziquone; 2,2',2''-trichlorotriethylamine; trichothecenes (especially T-2 toxin, verracurin A, roridin A and anguidine); urethan; vindesine; dacarbazine; mannomustine; mitobronitol; mitolactol; pipobroman; gacytosine; arabinoside ("Ara-C"); cyclophosphamide; taxoids, e.g., paclitaxel and docetaxel gemcitabine; 6-thioguanine; mercaptopurine; platinum coordination complexes such as cisplatin, oxaliplatin and carboplatin; vinblastine; platinum; etoposide (VP-16); ifosfamide; mitoxantrone; vincristine; vinorelbine; novantrone; teniposide; edatrexate; daunomycin; aminopterin; xeloda; ibandronate; irinotecan (e.g., CPT-11); topoisomerase inhibitor RFS 2000; difluorometlhylornithine (DMFO); retinoids such as retinoic acid; capecitabine; carboplatin, procarbazine, plicomycin, gemcitabien, navelbine, farnesyl-protein transferase inhibitors, transplatinum; and pharmaceutically acceptable salts, acids or derivatives of any of the above.

[00121] Immunotherapeutics, generally, rely on the use of immune effector cells and molecules to target and destroy cancer cells. The immune effector may be, for example, an antibody specific for some marker on the surface of a tumor cell. The antibody alone may serve as an effector of therapy or it may recruit other cells to actually effect cell killing. The antibody also may be conjugated to a drug or toxin (chemotherapeutic, radionuclide, ricin A chain, cholera toxin, pertussis toxin, etc.) and serve merely as a targeting agent. Alternatively, the effector may be a lymphocyte carrying a surface molecule that interacts, either directly or indirectly, with a tumor cell target. Various effector cells include cytotoxic T cells and NK cells as well as genetically engineered variants of these cell types modified to express chimeric antigen receptors.

[00122] The immunotherapy may comprise suppression of T regulatory cells (Tregs), myeloid derived suppressor cells (MDSCs) and cancer associated fibroblasts (CAFs). In some

embodiments, the immunotherapy is a tumor vaccine (e.g., whole tumor cell vaccines, peptides, and recombinant tumor associated antigen vaccines), or adoptive cellular therapies (ACT) (e.g., T cells, natural killer cells, TILs, and LAK cells). The T cells may be engineered with chimeric antigen receptors (CARs) or T cell receptors (TCRs) to specific tumor antigens. As used herein, a chimeric antigen receptor (or CAR) may refer to any engineered receptor specific for an antigen of interest that, when expressed in a T cell, confers the specificity of the CAR onto the T cell. Once created using standard molecular techniques, a T cell expressing a chimeric antigen receptor may be introduced into a patient, as with a technique such as adoptive cell transfer. In some aspects, the T cells are activated CD4 and/or CD8 T cells in the individual which are characterized by γ-IFN- producing CD4 and/or CD8 T cells and/or enhanced cytolytic activity relative to prior to the administration of the combination. The CD4 and/or CD8 T cells may exhibit increased release of cytokines selected from the group consisting of IFN-γ, TNF-α and interleukins. The CD4 and/or CD8 T cells can be effector memory T cells. In certain embodiments, the CD4 and/or CD8 effector memory T cells are characterized by having the expression of $CD44^{high}$ $CD62L^{low}$.

[00123] The immunotherapy may be a cancer vaccine comprising one or more cancer antigens, in particular a protein or an immunogenic fragment thereof, DNA or RNA encoding said cancer antigen, in particular a protein or an immunogenic fragment thereof, cancer cell lysates, and/or protein preparations from tumor cells. As used herein, a cancer antigen is an antigenic substance present in cancer cells. In principle, any protein produced in a cancer cell that has an abnormal structure due to mutation can act as a cancer antigen. In principle, cancer antigens can be products of mutated Oncogenes and tumor suppressor genes, products of other mutated genes, overexpressed or aberrantly expressed cellular proteins, cancer antigens produced by oncogenic viruses, oncofetal antigens, altered cell surface glycolipids and glycoproteins, or cell type-specific differentiation antigens. Examples of cancer antigens include the abnormal products of ras and p53 genes. Other examples include tissue differentiation antigens, mutant protein antigens, oncogenic viral antigens, cancer-testis antigens and vascular or stromal specific antigens. Tissue differentiation antigens are those that are specific to a certain type of tissue.

[00124] The immunotherapy may be an antibody, such as part of a polyclonal antibody preparation, or may be a monoclonal antibody. The antibody may be a humanized antibody, a chimeric antibody, an antibody fragment, a bispecific antibody or a single chain antibody. An antibody as disclosed herein includes an antibody fragment, such as, but not limited to, Fab, Fab' and F(ab')2, Fd, single-chain Fvs (scFv), single-chain antibodies, disulfide-linked Fvs (sdfv) and fragments including either a VL or VH domain. In some aspects, the antibody or fragment thereof specifically binds epidermal growth factor receptor (EGFR1, Erb-B1), HER2/neu (Erb-B2), CD20, Vascular endothelial growth factor (VEGF), insulin-like growth factor receptor (IGF-1R), TRAIL-receptor, epithelial cell adhesion molecule, carcino-embryonic antigen, Prostate-specific membrane antigen, Mucin-1, CD30, CD33, or CD40.

[00125] Examples of monoclonal antibodies include, without limitation, trastuzumab (anti-HER2/neu antibody); Pertuzumab (anti-HER2 mAb); cetuximab (chimeric monoclonal antibody to epidermal growth factor receptor EGFR); panitumumab (anti-EGFR antibody); nimotuzumab (anti-EGFR antibody); Zalutumumab (anti-EGFR mAb); Necitumumab (anti-EGFR mAb); MDX-210 (humanized anti-HER-2 bispecific antibody); MDX-210 (humanized anti-HER-2 bispecific antibody); MDX-447 (humanized anti-EGF receptor bispecific antibody); Rituximab (chimeric murine/human anti-CD20 mAb); Obinutuzumab (anti-CD20 mAb); Ofatumumab (anti-CD20 mAb); Tositumumab-I131 (anti-CD20 mAb); Ibritumomab tiuxetan (anti-CD20 mAb); Bevacizumab (anti-VEGF mAb); Ramucirumab (anti-VEGFR2 mAb); Ranibizumab (anti-VEGF mAb); Aflibercept (extracellular domains of VEGFR1 and VEGFR2 fused to IgG1 Fc); AMG386 (angiopoietin-1 and -2 binding peptide fused to IgG1 Fc); Dalotuzumab (anti-IGF-1R mAb); Gemtuzumab ozogamicin (anti-CD33 mAb); Alemtuzumab (anti-Campath-1/CD52 mAb); Brentuximab vedotin (anti-CD30 mAb); Catumaxomab (bispecific mAb that targets epithelial cell adhesion molecule and CD3); Naptumomab (anti-5T4 mAb); Girentuximab (anti-Carbonic anhydrase ix); or Farletuzumab (anti-folate receptor). Other examples include antibodies such as Panorex^{TM} (17-1A) (murine monoclonal antibody); Panorex (MAb17-1A) (chimeric murine monoclonal antibody); BEC2 (ami-idiotypic mAb, mimics the GD epitope) (with BCG); Oncolym (Lym-1 monoclonal antibody); SMART M195 Ab, humanized 13' 1 LYM-1 (Oncolym), Ovarex (B43.13, anti-idiotypic mouse mAb); 3622W94

mAb that binds to EGP40 (17-1A) pancarcinoma antigen on adenocarcinomas; Zenapax (SMART Anti-Tac (IL-2 receptor); SMART M195 Ab, humanized Ab, humanized); NovoMAb-G2 (pancarcinoma specific Ab); TNT (chimeric mAb to histone antigens); TNT (chimeric mAb to histone antigens); Gliomab-H (Monoclonals-Humanized Abs); GNI-250 Mab; EMD-72000 (chimeric-EGF antagonist); LymphoCide (humanized IL.L.2 antibody); and MDX-260 bispecific, targets GD-2, ANA Ab, SMART IDIO Ab, SMART ABL 364 Ab or ImmuRAIT-CEA. Further examples of antibodies include Zanulimumab (anti-CD4 mAb), Keliximab (anti-CD4 mAb); Ipilimumab (MDX-101; anti-CTLA-4 mAb); Tremilimumab (anti-CTLA-4 mAb); (Daclizumab (anti-CD25/IL-2R mAb); Basiliximab (anti-CD25/IL-2R mAb); MDX-1106 (anti-PD1 mAb); antibody to GITR; GC1008 (anti-TGF-β antibody); metelimumab/CAT-192 (anti-TGF-β antibody); lerdelimumab/CAT-152 (anti-TGF-β antibody); ID11 (anti-TGF-β antibody); Denosumab (anti-RANKL mAb); BMS-663513 (humanized anti-4-1BB mAb); SGN-40 (humanized anti-CD40 mAb); CP870,893 (human anti-CD40 mAb); Infliximab (chimeric anti-TNF mAb; Adalimumab (human anti-TNF mAb); Certolizumab (humanized Fab anti-TNF); Golimumab (anti-TNF); Etanercept (Extracellular domain of TNFR fused to IgG1 Fc); Belatacept (Extracellular domain of CTLA-4 fused to Fc); Abatacept (Extracellular domain of CTLA-4 fused to Fc); Belimumab (anti-B Lymphocyte stimulator); Muromonab-CD3 (anti-CD3 mAb); Otelixizumab (anti-CD3 mAb); Teplizumab (anti-CD3 mAb); Tocilizumab (anti-IL6R mAb); REGN88 (anti-IL6R mAb); Ustekinumab (anti-IL-12/23 mAb); Briakinumab (anti-IL-12/23 mAb); Natalizumab (anti-α4 integrin); Vedolizumab (anti-α4 β7 integrin mAb); T1 h (anti-CD6 mAb); Epratuzumab (anti-CD22 mAb); Efalizumab (anti-CD11a mAb); and Atacicept (extracellular domain of transmembrane activator and calcium-modulating ligand interactor fused with Fc).

[00126] When monitoring a mammal having, or suspected of having, cancer as described herein (e.g., based, at least in part, on differentially methylated CpG's in specific sequence contexts to identify differences in cfDNA fragmentation between healthy individuals and cancer patients), the monitoring can be before, during, and/or after the course of a cancer treatment. Methods of monitoring provided herein can be used to determine the efficacy of one or more cancer treatments and/or to select a mammal for increased monitoring. In some embodiments,

the monitoring can include identifying a cfDNA fragmentation profile as described herein. For example, a cfDNA fragmentation profile can be obtained before administering one or more cancer treatments to a mammal having, or suspected or having, cancer, one or more cancer treatments can be administered to the mammal, and one or more samples can be analyzed during the course of the cancer treatment. In some embodiments, a cfDNA fragmentation profile can change during the course of cancer treatment (e.g., any of the cancer treatments described herein). For example, a cfDNA fragmentation profile indicative that the mammal has cancer can change to a cfDNA fragmentation profile indicative that the mammal does not have cancer. Such a cfDNA fragmentation profile change can indicate that the cancer treatment is working. Conversely, a cfDNA fragmentation profile can remain static (e.g., the same or approximately the same) during the course of cancer treatment (e.g., any of the cancer treatments described herein). Such a static cfDNA fragmentation profile can indicate that the cancer treatment is not working.

[00127] In some embodiments, the monitoring can include conventional techniques capable of monitoring one or more cancer treatments (e.g., the efficacy of one or more cancer treatments). In some embodiments, a mammal selected for increased monitoring can be administered a diagnostic test (e.g., any of the diagnostic tests disclosed herein) at an increased frequency compared to a mammal that has not been selected for increased monitoring. For example, a mammal selected for increased monitoring can be administered a diagnostic test at a frequency of twice daily, daily, bi-weekly, weekly, bi-monthly, monthly, quarterly, semi-annually, annually, or any at frequency therein. In some embodiments, a mammal selected for increased monitoring can be administered a one or more additional diagnostic tests compared to a mammal that has not been selected for increased monitoring. For example, a mammal selected for increased monitoring can be administered two diagnostic tests, whereas a mammal that has not been selected for increased monitoring is administered only a single diagnostic test (or no diagnostic tests). In some embodiments, a mammal that has been selected for increased monitoring can also be selected for further diagnostic testing. Once the presence of a tumor or a cancer (e.g., a cancer cell) has been identified (e.g., by any of the variety of methods disclosed herein), it may be beneficial for the mammal to undergo both increased monitoring (e.g., to

assess the progression of the tumor or cancer in the mammal and/or to assess the development of one or more cancer biomarkers such as mutations), and further diagnostic testing (e.g., to determine the size and/or exact location (e.g., tissue of origin) of the tumor or the cancer). In some embodiments, one or more cancer treatments can be administered to the mammal that is selected for increased monitoring after a cancer biomarker is detected and/or after the differentially methylated CpG's in specific sequence contexts to identify differences in cfDNA fragmentation between healthy individuals and cancer patients of the mammal has not improved or deteriorated. Any of the cancer treatments disclosed herein or known in the art can be administered. For example, a mammal that has been selected for increased monitoring can be further monitored, and a cancer treatment can be administered if the presence of the cancer cell is maintained throughout the increased monitoring period. Additionally, or alternatively, a mammal that has been selected for increased monitoring can be administered a cancer treatment, and further monitored as the cancer treatment progresses. In some embodiments, after a mammal that has been selected for increased monitoring has been administered a cancer treatment, the increased monitoring will reveal one or more cancer biomarkers (e.g., mutations). In some embodiments, such one or more cancer biomarkers will provide cause to administer a different cancer treatment (e.g., a resistance mutation may arise in a cancer cell during the cancer treatment, which cancer cell harboring the resistance mutation is resistant to the original cancer treatment).

[00128] When a mammal is identified as having cancer as described herein (e.g., based, at least in part, on the cfDNA fragmentation profile of the mammal), the identifying can be before and/or during the course of a cancer treatment. Methods of identifying a mammal as having cancer provided herein can be used as a first diagnosis to identify the mammal (e.g., as having cancer before any course of treatment) and/or to select the mammal for further diagnostic testing. In some embodiments, once a mammal has been determined to have cancer, the mammal may be administered further tests and/or selected for further diagnostic testing. In some embodiments, methods provided herein can be used to select a mammal for further diagnostic testing at a time period prior to the time period when conventional techniques are capable of diagnosing the mammal with an early-stage cancer. For example, methods provided herein for selecting a

mammal for further diagnostic testing can be used when a mammal has not been diagnosed with cancer by conventional methods and/or when a mammal is not known to harbor a cancer. In some embodiments, a mammal selected for further diagnostic testing can be administered a diagnostic test (e.g., any of the diagnostic tests disclosed herein) at an increased frequency compared to a mammal that has not been selected for further diagnostic testing. For example, a mammal selected for further diagnostic testing can be administered a diagnostic test at a frequency of twice daily, daily, bi-weekly, weekly, bi-monthly, monthly, quarterly, semi-annually, annually, or any at frequency therein. In some embodiments, a mammal selected for further diagnostic testing can be administered a one or more additional diagnostic tests compared to a mammal that has not been selected for further diagnostic testing. For example, a mammal selected for further diagnostic testing can be administered two diagnostic tests, whereas a mammal that has not been selected for further diagnostic testing is administered only a single diagnostic test (or no diagnostic tests). In some embodiments, the diagnostic testing method can determine the presence of the same type of cancer (e.g., having the same tissue or origin) as the cancer that was originally detected (e.g., based, at least in part, on the cfDNA fragmentation profile of the mammal). Additionally, or alternatively, the diagnostic testing method can determine the presence of a different type of cancer as the cancer that was original detected. In some embodiments, the diagnostic testing method is a scan. In some embodiments, the scan is a computed tomography (CT), a CT angiography (CTA), an esophagram (a Barium swallow), a Barium enema, a magnetic resonance imaging (MRI), a PET scan, an ultrasound (e.g., an endobronchial ultrasound, an endoscopic ultrasound), an X-ray, a DEXA scan.

[00129] In some embodiments, the diagnostic testing method is a physical examination, such as an anoscopy, a bronchoscopy (e.g., an autofluorescence bronchoscopy, a white-light bronchoscopy, a navigational bronchoscopy), a colonoscopy, a digital breast tomosynthesis, an endoscopic retrograde cholangiopancreatography (ERCP), an ensophagogastroduodenoscopy, a mammography, a Pap smear, a pelvic exam, a positron emission tomography and computed tomography (PET-CT) scan. In some embodiments, a mammal that has been selected for further diagnostic testing can also be selected for increased monitoring. Once the presence of a tumor or a cancer (e.g., a cancer cell) has been identified (e.g., by any of the variety of methods disclosed

herein), it may be beneficial for the mammal to undergo both increased monitoring (e.g., to assess the progression of the tumor or cancer in the mammal and/or to assess the development of one or more cancer biomarkers such as mutations), and further diagnostic testing (e.g., to determine the size and/or exact location of the tumor or the cancer). In some embodiments, a cancer treatment is administered to the mammal that is selected for further diagnostic testing after a cancer biomarker is detected and/or after the cfDNA fragmentation profile of the mammal has not improved or deteriorated. Any of the cancer treatments disclosed herein or known in the art can be administered. For example, a mammal that has been selected for further diagnostic testing can be administered a further diagnostic test, and a cancer treatment can be administered if the presence of the tumor or the cancer is confirmed. Additionally, or alternatively, a mammal that has been selected for further diagnostic testing can be administered a cancer treatment and can be further monitored as the cancer treatment progresses. In some embodiments, after a mammal that has been selected for further diagnostic testing has been administered a cancer treatment, the additional testing will reveal one or more cancer biomarkers. In some embodiments, such one or more cancer biomarkers (e.g., mutations) will provide cause to administer a different cancer treatment (e.g., a resistance mutation may arise in a cancer cell during the cancer treatment, which cancer cell harboring the resistance mutation is resistant to the original cancer treatment).

[00130]

[00131] Examples

[00132] Example 1: DNA Methylation and Gene Expression as Determinants of Genome-Wide Cell-Free DNA Fragmentation.

[00133] With the development of high-throughput sequencing methods, it has become possible to study features of cfDNA fragmentation, including those related to the underlying nucleosomes. In healthy individuals, the positioning of nucleosomes as well as larger chromatin compartments shows striking similarity to those of myelocytic and lymphocytic cells[8–10]. Similarly, methylation profiles of cfDNA from individuals without cancer are very similar to DNA methylation of leukocytes[11]. Although epigenetic changes are related to genome packaging[12] and chromatin structure[13] as well as gene expression[14], until now there have been

limited studies of the connection between methylation, expression and cfDNA fragmentation[15–17]. Additionally, none of these studies have examined the underlying impact of these changes on recurrent cfDNA breakpoint motifs and fragment size.

[00134] In this study the features related to motifs associated with nucleosome positioning and breakpoints of cfDNA fragments were analyzed in both healthy individuals and patients with cancer. It was shown how epigenetic marks give rise to specific patterns of fragmentation of cfDNA and how these are related to both methylation and gene expression. Using this information, it was shown that differentially methylated CpG's in specific sequence contexts can be used to identify differences in cfDNA fragmentation between healthy individuals and cancer patients.

[00135] **METHODS**

[00136] *Study Populations*

[00137] For analyzing motif frequencies, recurrent ends and the relationship with gene expression and methylation, low coverage whole genome sequencing (WGS) of cfDNA (1-2x) from 787 individuals without cancer as well as 182 individuals with cancer was used[32]. cfDNA methylation from individuals without cancer was previously analyzed using Illumina's Infinium methylationEPIC array[11] and made available through NCBI's Gene Expression Omnibus (GEO) database (dataset identifier GSE122126). The cells that contribute to cfDNA in individuals without cancer, was used and validated previously[9,11], showing most cfDNA originates from myeloid-derived cells. Average gene expression was used from multiple myeloid cell lines, as previously published[33].

[00138] *Processing of cfDNA Samples*

[00139] Whole-genome libraries of cancer patients and cancer-free individuals were sequenced using 100-bp paired-end runs (200 cycles) on the Illumina HiSeq2500 platform at 1–2x coverage per genome. Prior to alignment, adapter sequences were filtered from reads using the fastp software[34]. Sequence reads were aligned against the hg19 human reference genome using Bowtie235 and duplicate reads were removed using Sambamba[36]. Only reads with a mapq score of at least 30 or greater were retained. Post-alignment, each aligned pair was converted to a genomic interval representing the sequenced DNA fragment using bedtools[37].

[00140] *Frequency of motifs around the ends of cfDNA fragments*

[00141] The expected frequency of 3bp motifs in the human genome was calculated by counting the occurrence of each 3bp motif in the human genome (hg19). For computing the empirical frequency of 3bp motifs at the ends of sheared (sonicated) fragments, we used published sequencing data from 10 lymphoblastoid cell lines[38]. The genomic DNA from these lymphoblastoid cell lines was fragmented through sonication with a Covaris M220 Focused Ultrasonicator. For this analysis, the data was re-analyzed on the same way as cfDNA samples were analyzed. For the 10 lymphoblastoid cell lines and the 543 low coverage WGS of individuals without cancer, the number of 3bp motifs were counted at the start of the fragment and the reverse complement of the 3bp motif around the end of the fragment. The 3bp motif contains 1 base outside the fragment, followed by the first 2 bases of the fragment. Using these absolute numbers, relative frequencies of each of the 64 3bp motifs were calculated.

[00142] To quantify the preference for cfDNA fragments to end at a specific genomic location, the ratio of cfDNA fragments ending at this location (recurrent fragment ends) divided by the number of cfDNA fragments having a 1bp or more overlap within +/- 50bps of this position (neighboring fragments) were calculated. This ratio of recurrent fragment ends to neighboring fragments was computed by aggregating cfDNA fragments across all 543 individuals without cancer. This calculation was repeated for every position in the hg19 reference genome where the number of neighboring fragments was comprised of 200 or more high quality alignments (mapq > 30), thereby capturing recurrence in non-repetitive regions of the genome. Genomic positions with high recurrence were defined as those where the ratio of recurrent fragment ends to neighboring fragments was at least 5%.

[00143] *X-Ray Crystal Structure of Nucleosomes*

[00144] To identify structures of the nucleosome bound to DNA, the Protein Data Bank (PDB; rcsb.org) was searched for the term "Nucleosome" and the results were filtered for those structures derived from x-ray diffraction or cryo-EM leading to 427 entries. DNA sequences from these structures (648) were downloaded and filtered for sequences that were at least 167 in length. This identified 80 unique sequences from 51 PDB entries. These entries were visually inspected and those with less than 167 bases resolved or where the interaction with the H1 linker

was disrupted by another DNA binding protein were removed. This left 17 structures (Table 3). Motifs were considered well positioned if they were within 5 angstroms of the H1 linker or if the bases 167 away on the same strand were within 5 angstroms of the H1 linker.

[00145] *Connecting cfDNA Fragment Patterns to CpG Methylation*

[00146] In order to discover whether fragmentation patterns in cfDNA from individuals without cancer were influenced by methylation, raw data was analyzed from Illumina's Infinium methylationEPIC array from 8 different cfDNA experiments, with 4 biologically different cohorts (young men, old men, young women, old women)[11]. Standard pipelines were used to process the Infinium arrays[39]. A numeric score (beta value) ranging between 0 (unmethylated) and 1 (methylated) was obtained at each CpG and averaged across the samples to summarize the overall level of methylation. CpG sites were labeled as unmethylated if the mean beta-value was < 0.3 and methylated if the mean beta-value was > 0.7. For 543 cfDNA plasma samples of individuals without cancer (Table 1), the position of Infinium CpG sites within the cfDNA fragments was recorded using a 1-base index. CpGs were grouped according to the mean beta beta- value from the Infinium arrays. The number of fragments were counted starting or ending at the CpG sites within the CpG group and this frequency was scaled by the number of fragments having any overlap within a 50bp window of the start- or end-position. Fragments were further categorized according to their 3bp end motif and whether the cfDNA fragment was located in a CpG-island, shore, shelf, or open sea.

[00147] *cfDNA Sequence Coverage and Fragment Sizes at CpG-islands and Transcription Start Sites (TSSs).*

[00148] To summarize cfDNA fragment lengths at one CpG island, the average length of cfDNA fragments was counted starting or ending at the CpG island. By convention, this was referred to as position 0. This summarization step was repeated at positions x bps from the CpG island in 1 bp increments ranging from x = -3000bp to x = +3000bp as well as x = -500000bp to x = +500000bp. This procedure was repeated for each of the CpG islands. Mean fragment lengths at TSSs were summarized in a similar manner with position 0 denoting the TSS. As low coverage WGS was used, cfDNA fragments were pooled across all 543 non-cancer samples. Using the previously described methylation data[11] and gene-expression data from myeloid cell

lines[33], the regions were ordered so as to visualize patterns that were associated with CpG-island methylation and gene expression (FIGS. 3A, 3B, 3C, 3D).

[00149] *Gene Set Enrichment Analyses*

[00150] Gene set enrichment analysis[40] was performed with the Hallmark[41] and KEGG[42] gene sets acquired from the Molecular Signatures Database[43]. Averaged normalized counts across healthy PBMC samples were used for ranking genes in RNA sequencing. Averaged beta-values for all CpGs overlapping transcripts were used for ranking genes in the methylation analyses. These metrics are identical to those used previously to rank genes for visualization of cfDNA coverage surrounding CpG islands and transcription start sites. For coverage based analyses, the normalized fragment count overlapping the transcription start site or the CpG islands falling within transcripts was used. 10,000 permutations were run for each set of ranks, leading to a minimum unadjusted p-value of 1e-4. All gene sets that were moderately significant across all comparisons in any direction (unadjusted $p < .1$) were selected for inclusion in the heatmap showing enrichment scores by gene set (FIG. 3G).

[00151] *Multivariate model*

[00152] Generalized linear models were used to evaluate the relationship between the aggregated mean cfDNA fragment size and total coverage at the transcript level with RNA expression, WPS, and methylation. For methylation, we calculated the mean beta-value at each CpG-island across 97 blood samples processed on the Infinium array (see *Study populations*). CpG-islands were mapped to transcripts by their proximity to TSSs using the R-package annotatr (version 1.28.0). A transcript was considered methylated if the mean beta value was 0.5 or higher and unmethylated otherwise. The mean RNA expression (mean TPM) across 6 myeloid cell lines were transformed as log10(mean TPM + 1) and then centered and scaled by the overall mean and standard deviation across all transcripts, respectively. WPS was summarized for each transcript in the interval +1 to +10 bases from the TSS and centered and scaled. Total cfDNA coverage across 543 non-cancers was calculated at each base in the interval -10bp to -1bp from the TSS and averaged, while mean fragment sizes were calculated in the interval from -1480bp to -1471bp from the TSS. The intervals for summarizing cfDNA coverage, fragment size, and WPS were evaluated for all 10bp genomic intervals within 2500bp from the TSS. The interval

that yielded measurements with maximum absolute correlation to RNA expression was selected for the regression analyses. With these quantitative summaries as described above, the expected normalized coverage, DV, for transcript *i* is given by

[00153] $$E(DV_i|WPS_i, RNA_i, Meth_i) = \beta_0 + \beta_1 WPS_i + \beta_2 WPS_i^2 + \beta_3 WPS_i^3 + \beta_4 RNA_i + \beta_5 Meth_i + \beta_6 RNA_i \times Meth_i.$$

[00154] The expected fragment length was modeled in a similar fashion. Coefficients from these models were estimated using a generalized linear model with identity link function in R (version 4.3.2). Using analysis of variance (ANOVA), we assessed whether RNA expression helps explain variation in coverage after adjusting for methylation and WPS by testing both the main effect for RNA expression and its interaction with methylation. We performed a similar ANOVA to evaluate whether methylation explained variation in coverage after adjusting for the effects of RNA and WPS on coverage. Forest plots were generated for each model to visualize estimated model coefficients with 95% confidence intervals using sjPlot (version 2.8.15).

[00155] *Monte Carlo simulation on human cfDNA coverage in xenograft models*

[00156] Coverage was calculated for the top 500, 1000, 2000, 3000, 4000 and 5000 most differentially methylated CpG-islands or most differentially expressed genes for each of the six xenografts (3 IDH1 R132H mutant xenografts and 3 IDH1 wild-type xenografts). These coverages were normalized for the total size of these regions. Comparing IDH1 mutant to wild-type, we determined whether the direction of the difference in normalized coverage agreed with our a priori expectation that we would observe higher coverage in methylated regions and lower coverage in expressed regions for each of the four possible comparisons (high methylation regions in IDH mutant, high methylation in IDH wild-type, high expression in IDH mutant, high expression in IDH wild-type). To evaluate how likely we would observe the empirical agreement under the null hypothesis that there is no difference in cfDNA coverage in mice between IDH1 mutant and wild-type tumors, we permuted the mutant and wild-type labels and evaluated the agreement as previously described. This was repeated for 10,000 iterations. We repeated this process for each of the 19 possible permutations of the sample labels, deriving a distribution of agreement under the null. The p-value was computed as the proportion of permutations where the agreement was as high or higher than the empirical agreement obtained

from the non-permuted class labels. These analyses were repeated for each of the six region or gene list sizes indicated above for a total of 24 comparisons.

[00157] *Differentially Methylated CpG-Based Tumor Specific cfDNA Methylation Patterns*

[00158] Using publicly available data, differentially methylated CpGs data from individuals with and without cancer was evaluated. A large cohort of differentially methylated regions was published for pancreatic cancers[30]. Using these differentially methylated CpGs subgroups were made based on the direction of differential methylation (non-cancer methylated vs. pancreatic cancer unmethylated; non-cancer unmethylated vs. pancreatic cancer methylated) and based on the 3bp and 4bp motifs. In total, 16 different features were extracted for each sample, based on the ratio of aggregated cfDNA fragments starting or ending at these motifs, divided by the aggregated amount of fragments overlapping a 101bp-window around the motif (FIG. 4A). These features was used to generate a machine learning model, for which the performance was compared with the original DELFI-method8. An ensemble model, combining both the methylation-based approach and DELFI showed a synergistic increase in performance compared to the two models separately (FIG. 4B).

[00159] *Statistical Methods*

[00160] All t-tests were Welch two sample t-tests unless otherwise indicated. Statistical analyses were performed using R, version 4.2.0.

[00161] **RESULTS**

[00162] The frequency and composition of cfDNA start and end sequences were investigated, as these have been previously described as non-random and potentially related to cleavage by endogenous DNAses1[8,19]. To rigorously identify cfDNA end positions, cfDNA sequence data was pooled from low coverage whole genome sequencing from a cohort of healthy individuals (n=543) and the frequency of cfDNA breakpoints was investigated at every possible position in the genome. Only fragment reads with high sequence and mapping quality were considered and the ratio of the number of cfDNA fragments starting or ending at a particular position compared to the number of fragments with start or end positions within 50 bp surrounding that location, was calculated. This approach was necessary to identify recurrent

fragment end sequences and to account for differences in cfDNA fragment size and coverage across the genome[8].

[00163] cfDNA fragments with more frequently observed end-positions were evaluated and found that these were enriched for specific motifs. These typically included a thymine or an adenine before the start of the cfDNA fragment and two cytosines (A/T|CC) or a cytosine followed by a guanine (A/T|CG) as the first two nucleotides of the cfDNA fragment (FIG. 1A). It was reasoned that end sequences that occurred recurrently among cfDNA fragments would likely represent those locations protected by nucleosome occupancy and found that the frequency of these base motifs increased further at "preferred" recurrent ends among healthy individuals (FIGS. 1A, 1B). The occurrence of A/T|CC and A/T|CG preferred cfDNA fragment ends were observed at a rate much higher than theoretically expected in the genome (26.5x for A/T|CC and 5.5x for A/T|CG) (FIG. 1A)($p<2.2e-16$, one-sample t-test), while the frequencies of DNA ends from fragments generated through sonication of genomic DNA were close to theoretical abundances.

[00164] To understand the underlying basis for the enrichment of A/T|CG end sequences, available x-ray crystal and cryo-EM structures of DNA bound to nucleosomes was examined. It was found that in 82% of unique structures an A/T|CG motif was found close to the H1 linker or centered 167 bp away, the size of typical cfDNA molecules (FIG. 1C). In contrast to fragment end sequences, it was observed that interior regions of cfDNA fragments were enriched in adenines and thymines, with a 10-11bps periodicity in the frequency of these nucleotides over the length of the fragment (FIG. 5). These observations were consistent with current predictions of DNA wrapping around a histone-core and the necessity to alternate rigid DNA regions (C and G-rich) with more flexible regions (A- and T-rich) to wrap nearly two turns around the nucleosome[20].

[00165] Given the preponderance of fragment ends containing CGs, it was questioned whether epigenetic marks of these sites could affect cfDNA fragmentation. cfDNA fragments were mapped to the genome and their ends were evaluated with respect to previously identified methylated and unmethylated CpG sites of cfDNA from healthy individuals obtained from methylation arrays evaluating 850K CpG sites (Moss J. *et al. Nat Commun* 9, 5068 (2018)). It

was observed that methylated CpGs were enriched at the ends of A/T|CG cfDNA fragments, while unmethylated CpGs had relatively even distributions over the length of these molecules (FIGS. 2A, 7A, 7B). To quantitatively assess the enrichment of fragment ends at CpGs, for each CpG the fraction of cfDNA fragments was calculated starting or ending at this dinucleotide position over the number of fragments with start or end positions within 50 bp around each CpG. It was observed that the fraction of preferred ends increased as much as 2.4 fold with higher levels of methylation ($p<2.2e-16$, t-test) (FIG. 2B). Furthermore, CG cfDNA fragment ends were enriched as much as 2.2-fold at locations of methylated CpG's throughout the genome, including in CpG islands, shores, shelves and open sea regions (FIG. 6), revealing that enrichment of methylated CpG fragments was a universal characteristic of cfDNA in these regions.

[00166] It was observed that methylated CG end sequences were preferentially enriched even when they overlapped frequently observed CC fragment end sequences. When N|CC sequences were followed by guanine resulting in N|CCG, the typical N|CC end motifs were reduced in frequency as these competed with the overlapping C|CG motif that was enriched when CpG sites containing these sequences were methylated (FIGS. 2B, 2C). The overall impact of this competition resulted in a dramatic reduction of N|CCG fragment end sequences at methylated CpG positions that was even greater than the corresponding increase in fragment ends at N|CG, as seen for example with the 3.7 fold reduction in T|CCG end sequences ($p<2.2e-16$, t-test), compared to a 2.2 fold increase in C|CG end motifs ($p<2.2e-16$, t-test).

[00167] To provide additional biological evidence for the link between methylated CGs and cfDNA fragmentation, cfDNA fragments arising from the X chromosome was compared among healthy individuals, as it is well established that one copy of the two X chromosomes is inactivated by methylation of CpG islands in women, while these regions on the single X chromosome in men are not methylated[21,22]. In line with the observation of methylation-induced fragment end enrichment, cfDNA fragments ending with CG were enriched, and fragment ends with CCG were preferentially reduced, at locations of X chromosome CpG islands in women compared to men (for CG ends, average 0.39% men vs 0.48% women, $p<2.2e-16$, t-test; for CCG ends, average 0.68% men vs 0.54% women, $p<2.2e-16$, t-test), but these differences were not observed on the autosomes in men and women ($p=0.58$ and $p=0.80$, respectively, t-test)(FIG.

2D). Although this trend continued in CpG shores, higher CG fragment end enrichment was observed in men compared to women in CpG shelves and open sea, consistent with the previously reported increased methylation on the male X chromosome in these regions (FIG. 8)[12,23].

[00168] In addition to the enrichment of cfDNA fragment-end positions at sites of epigenetic marks, it was observed that cfDNA sequence coverage (the average number of cfDNA molecules overlapping a specific position) was related to methylation levels ($r=0.6$, $p<2.2e-16$, Pearson correlation test) (FIG. 9A), and was up to 1.7 fold higher across regions of CpG islands that were methylated compared to those that were not methylated ($p<2.2e-16$, t-test) (FIG. 3A). Given the connection between CpG island methylation and expression, the relationship between gene expression at transcription start sites (TSS) and cfDNA fragmentation patterns was evaluated. There was an inverse relationship between cfDNA coverage at TSS and expression levels of nearby genes ($r=-0.48$, $p<2.2e-16$, Pearson correlation test) (FIG. 9B). Overall levels of cfDNA fragments that overlapped TSSs of unexpressed genes were up to 3.7-fold higher than at regions of expressed genes, likely due to the lack of destabilizing effects of transcription factors on nucleosomes ($p<2.2e-16$, t-test) (FIG. 3B). Concordant with higher cfDNA coverage, changes in cfDNA fragment sizes at these regions was observed, including fragments 4-5 bp smaller at areas 800-1000 bp upstream of TSSs of highly expressed compared to unexpressed genes (164.5 bp vs 168.6 bp, respectively, $p<2.2e-16$, t-test) or in regions surrounding unmethylated CpG islands compared to highly methylated CpGs (165.1 bp vs 167.3 bp, respectively, $p<2.2e-16$, t-test) (FIGS. 3C, 3D). Examining broader regions surrounding TSS or CpG islands continued to reveal differences between expressed/unexpressed or unmethylated/methylated genes in regions as far as 500 kb around these sites (FIGS. 10A-10D).

[00169] An analysis of cfDNA fragments adjacent to genes in KEGG and Hallmark gene sets revealed that cfDNA coverage was concordant with expression and CpG methylation[11] across all significant gene sets identified in blood cells ($p<0.1$, gene set enrichment analysis) (FIG. 4A, Table 7). These included higher cfDNA coverage at regions of CpG islands and TSSs when methylation was increased, and expression was decreased, and lower cfDNA coverage with decreased methylation and increased expression. For example, gene pathways not typically

expressed in blood cells, including neuronal receptor-ligand interactions or olfactory receptor transduction, were typically methylated and more highly represented in cfDNA fragments at regions containing CpG islands or TSSs (FIG. 3F). In contrast, genes utilized in hematopoiesis, including in E2f transcription factor targets and blood cell metabolism genes were highly expressed, more frequently unmethylated, and represented at lower cfDNA levels at CpG or TSS regions of these genes (FIG. 4B). Overall, it was found that cfDNA coverage was related to both CpG methylation and expression of nearby genes (FIG. 11A), but that recurrent cfDNA fragment end enrichment at CpG sites was more closely related to methylation levels than gene expression (FIG. 11B). These results highlight that DNA methylation is a fundamental feature affecting cfDNA fragmentation that is associated with but independent of gene expression.

[00170] To provide a direct and independent analysis of the effect of methylation or gene expression with cfDNA coverage, we assessed human cfDNA fragmentation coverage in the plasma of mice with implanted human tumors with or without a knock-in of the IDH1 chromatin modifier with a mutation at R132 that was known to be activating through our previous work (Parsons, D. W. *et al. Science* **321**, 1807–1812 (2008)) and lead to widespread genome-wide methylation and expression changes. Duncan, C. G. *et al. Genome Res.* **22**, 2339–2355 (2012).; Wei, S. *et al. Oncogene* **37**, 5160–5174 (2018); Brennan, C. W. *et al. Cell* **155**, 462–477 (2013). Mice were injected with U87 glioblastoma cell line that was wild-type for IDH1 (n=3) or isogenically altered to contain the R132H mutation (n=3) and evaluated at 20-30 days after tumor implantation. After selecting for human-derived cfDNA fragments from the mouse plasma, high coverage of human cfDNA was observed at areas of increased methylation, while low coverage of cfDNA was present at regions of increased expression (FIG. 3G) ($p<0.053$, Monte Carlo simulation), consistent with our previous analyses. This well-controlled analysis provides a direct causal link between genome-wide changes in epigenetic features and cfDNA fragmentation.

[00171] As it has been widely reported that the overall size of cfDNA is smaller in patients with cancer compared to that of healthy individuals[24,25], it was speculated whether genome-wide changes in DNA methylation and gene expression during tumorigenesis[26] may have an effect on cfDNA fragmentation in cancer patients. To unambiguously compare tumor-

derived with wild-type cfDNA, changes in cfDNA fragment sizes of tumor derived and WBC derived cfDNA were examined using ultrasensitive NGS targeted sequencing from 98 patients with cancer[8,27]. An average shift of 3.9 bps was found in tumor-derived cfDNA of these patients that was similar to the observed cfDNA size differences at TSS regions of high and low expression and CpG sites in methylated versus unmethylated regions of healthy cfDNA (FIG. 3E). As we observed that tumors typically have an increased number of expressed genes and are hypomethylated compared to white blood cells (FIGS. 12A, 12B), similar to previous studies26,28,29 these results support the notion that changes in expression and methylation in cancer cells may in part be responsible for the overall smaller cfDNA fragments observed in patients with cancer.

[00172] To identify the impact of differences in CpG methylation on cfDNA fragment ends between healthy individuals and patients with cancer, regions previously identified by comparison of reduced representation bisulfite sequencing (RRBS) data from laser capture microdissected pancreatic ductal adenocarcinoma and normal pancreatic tissues were evaluated, and where these regions were confirmed in cfDNA using methyl-DNA immunoprecipitation (MeDIP)[30]. The fragment end representation at CG and CCG sites was then assessed through low coverage whole genome cfDNA analyses of patients with pancreatic (n=34), colorectal (n=27), ovarian (n=28), lung (n=39) or breast cancer (n=54) as well as from individuals without cancer (n=244)[8]. In regions with increased CpG methylation in non-cancer tissues, a preferential decrease was observed in cfDNA fragments ending with N|CCG in individuals without cancer compared to the abundance of these fragments in patients with pancreatic and other cancers (FIG. 5A). In contrast, in regions with increased methylation in pancreatic cancer, an increase in cfDNA fragments was observed ending with CG in patients with cancer compared to levels in individuals without cancer (FIG. 5A). In all cases the strongest signal was observed in the patients with pancreatic cancer, suggesting that the use of tumor-specific sites of methylation resulted in improved performance in this tumor type. Incorporation of the distribution of fragment end positions at these CG and CCG sites in a gradient boosted tree machine learning model successfully distinguished individuals with pancreatic cancer compared to those without (cross-validated AUC=0.87). Combining this approach with genome-wide fragmentation

analyses (DELFI)8 that incorporate fragment coverage and size improved the sensitivity of the combined method (AUC=0.94, 95% CI=0.90-0.96). These observations suggested that genome-wide DNA methylation can be used to detect individuals with cancer using cfDNA fragment end representation at CG and CCG sites.

[00173] A multivariate regression model evaluating DNA methylation (FIG. 3A, 3C), gene expression (FIG. 3B, 3D), nucleosome positioning (Supplementary Fig. 13), and the interaction of these terms revealed that each of these elements contributed independently to cfDNA coverage and fragment size (FIG. 3F). The relationship between methylation and coverage was qualitatively similar in more complex models that included additional terms for the interaction of DNA methylation and nucleosome positioning, and the three-way interaction of DNA methylation, gene expression, and nucleosome positioning (FIG. 13). These results highlight that DNA methylation is a fundamental feature affecting cfDNA fragmentation.

[00174] **DISCUSSION**

[00175] In this study, it was shown inter alia that genome wide CpG methylation has a profound impact on cfDNA fragmentation. Recurrent cfDNA sequences with CG fragment ends were enriched at sites of methylation, increasing at N|CG and decreasing at competing N|CCG sites in a manner that was dependent on the level of methylation. Structural analyses of DNA bound to nucleosomes showed that CG sequences were typically located close to the histone H1 linker. These observations, together with previous molecular dynamic simulations[31], evidence that methylation of CG sequences may provide a more stable interaction between the methylated DNA and the H1 linker, thereby protecting nucleosome bound cfDNA fragments from degradation.

[00176] Methylated CpGs affected not only fragment end positions, but also resulted in a higher amounts of circulating cfDNA at these regions. cfDNA fragmentation was similarly affected at TSS of genes with decreased expression, both at the individual gene level, as well as in gene pathways. cfDNA fragment sizes were altered by both methylation and expression changes, and could be observed nearby CG and TSS sites, as well as at distances hundreds of thousands of bases away. These determinants of cfDNA size, coupled with an overall increase in gene expression and decrease in methylation in human cancers observed in this study, provide

evidence of a mechanism for the global reduction of cfDNA fragment lengths observed in cancer patients.

[00177] Incorporation of cfDNA fragment end features at CpG sites into a cross-validated machine learning model provides an approach that can be used to detect cancer independently of other cfDNA characteristics. This approach appeared complementary to the DELFI cfDNA fragmentation analyses and together resulted in a method that had improved performance. Understanding methylation and expression differences affecting cfDNA size and coverage in patients with cancer will improve methods for assessing genome-wide cfDNA fragmentation in the future. Integration of methylation and expression changes with other genome-wide epigenetic marks provides complementary insights into the origins and mechanisms of cfDNA fragmentation.

[00178] **References**

1. Richmond, T. J., Finch, J. T., Rushton, B., Rhodes, D. & Klug, A. Structure of the nucleosome core particle at 7 Å resolution. *Nature* 311, 532–537 (1984).

2. Richmond, T. J. & Davey, C. A. The structure of DNA in the nucleosome core. *Nature* 423, 145–150 (2003).

3. Ceppellini, R., Polli, E. & Celada, F. A DNA-Reacting Factor in Serum of a Patient with Lupus Erythematosus Diffusus.∗. *P Soc Exp Biol Med* 96, 572–574 (1957).

4. Robbins, W. C., Holman, H. R., Deicher, H. & Kunkel, H. G. Complement Fixation with Cell Nuclei and DNA in Lupus Erythematosus. *P Soc Exp Biol Med* 96, 575–579 (1957).

5. Miescher, P. & Strässle, R. New Serological Methods for the Detection of the L. E. Factor*. *Vox Sang* 2, 283–287 (1957).

6. Seligmann, M. [Demonstration in the blood of patients with disseminated lupus erythematosus a substance determining a precipitation reaction with desoxyribonucleic acid]. *Comptes Rendus Hebdomadaires Des Seances De L'academie Des Sci* 245, 243–5 (1957).

7. Barra, G. B. *et al.* EDTA-mediated inhibition of DNases protects circulating cell-free DNA from ex vivo degradation in blood samples. *Clin Biochem* 48, 976–981 (2015).

8. Cristiano, S. *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature.* 570, 385–389 (2019).

9. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* 164, 57–68 (2016).

10. Foda, Z. H. *et al.* Detecting liver cancer using cell-free DNA fragmentomes. *Cancer Discov* 13, 616–631 (2022).

11. Moss, J. *et al.* Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat Commun* 9, 5068 (2018).

12. Fortin, J.-P. & Hansen, K. D. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol* 16, 180 (2015).

13. Collings, C. K., Waddell, P. J. & Anderson, J. N. Effects of DNA methylation on nucleosome stability. *Nucleic Acids Res* 41, 2918–2931 (2013).

14. Keshet, I., Yisraeli, J. & Cedar, H. Effect of regional DNA methylation on gene expression. *Proc National Acad Sci* 82, 2560–2564 (1985).

15. Ulz, P. *et al.* Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nat Genet* 48, 1273–1278 (2016).

16. Esfahani, M. S. *et al.* Inferring gene expression from cell-free DNA fragmentation profiles. *Nat Biotechnol* 40, 585–597 (2022).

17. Zhou, Q. *et al.* Epigenetic analysis of cell-free DNA by fragmentomic profiling. *Proc National Acad Sci* 119, e2209852119 (2022).

18. Chan, K. C. A. *et al.* Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. *Proc National Acad Sci* 113, E8159–E8168 (2016).

19. Serpas, L. *et al.* Dnase1l3 deletion causes aberrations in length and end-motif frequencies in plasma DNA. *Proc National Acad Sci* 116, 641–649 (2019).

20. Trifonov, E. N. Cracking the chromatin code: Precise rule of nucleosome positioning. *Phys Life Rev* 8, 39–50 (2011).

21. Norris, D. P., Brockdorff, N. & Rastan, S. Methylation status of CpG-rich islands on active and inactive mouse X chromosomes. *Mamm Genome* 1, 78–83 (1991).

22. Tribioli, C. *et al.* Methylation and sequence analysis around Eagi sites: identification of 28 new CpG islands in XQ24-XQ28. *Nucleic Acids Res* 20, 727–733 (1992).

23. Duncan, C. G. *et al.* Dosage compensation and DNA methylation landscape of the X chromosome in mouse liver. *Sci Rep-uk* 8, 10138 (2018).

24. Jiang, P. *et al.* Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc National Acad Sci* 112, E1317–E1325 (2015).

25. Giacona, M. B. *et al.* Cell-Free DNA in Human Blood Plasma. *Pancreas* 17, 89–97 (1998).

26. Jones, P. A. & Baylin, S. B. The fundamental role of epigenetic events in cancer. *Nat Rev Genet* 3, 415–428 (2002).

27. Phallen, J. *et al.* Direct detection of early-stage cancers using circulating tumor DNA. *Sci Transl Med* 9, eaan2415 (2017).

28. Baylin, S. B. *et al.* Abnormal patterns of DNA methylation in human neoplasia: potential consequences for tumor progression. *Cancer Cells Cold Spring Harb N Y 1989* 3, 383–90 (1991).

29. Gama-Sosa, M. A. *et al.* The 5-methylcytosine content of DNA from human tumors. *Nucleic Acids Res* 11, 6883–6894 (1983).

30. Shen, S. Y. *et al.* Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* 563, 579–583 (2018).

31. Li, S., Peng, Y., Landsman, D. & Panchenko, A. R. DNA methylation cues in nucleosome geometry, stability and unwrapping. *Nucleic Acids Res* 50, 1864–1874 (2022).

32. Mathios, D. *et al.* Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat Commun* 12, 5060 (2021).

33. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* 347, 1260419 (2015).

34. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor.

Bioinformatics 34, i884–i890 (2018).

35. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2.

Nat Methods 9, 357–359 (2012).

36. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 31, 2032–2034 (2015).

37. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842 (2010).

38. Papp, E. *et al.* Integrated Genomic, Epigenomic, and Expression Analyses of Ovarian Cancer Cell Lines. *Cell Reports* 25, 2617–2633 (2018).

39. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30, 1363–1369 (2014).

40. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc National Acad Sci* 102, 15545–15550 (2005).

41. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst* 1, 417–425 (2015).

42. Kanehisa, M. & Goto, S. Kegg: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 28, 27–30 (2000).

43. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740 (2011).

44. Adhireksan, Z. *et al.* Engineering nucleosomes for generating diverse chromatin assemblies. *Nucleic Acids Res* 49, gkab070 (2021).

OTHER EMBODIMENTS

[00179] From the foregoing description, it will be apparent that variations and modifications may be made to the disclosure described herein to adopt it to various usages and conditions. Such embodiments are also within the scope of the following claims.

[00180] All citations to sequences, patents and publications in this specification are herein incorporated by reference to the same extent as if each independent patent and publication was specifically and individually indicated to be incorporated by reference. By their citation of various references in this document, Applicants do not admit any particular reference is "prior art" to their disclosure.

What is claimed:

1. A method for determining circulating cell-free DNA (cfDNA) fragmentation in a sample, comprising:

assaying cfDNA sequences to identify the genomic location and end positions of cfDNA fragments;

analyzing frequency of cfDNA breakpoints at a plurality of positions in the genomic sequences; thereby,

determining circulating cell-free DNA (cfDNA) fragmentation.

2. The method of claim 1, wherein the genomic sequences are assayed by whole genome sequencing or obtaining whole genome sequences from a database and pooling cfDNA sequences.

3. The method of claim 1, wherein analysis of the frequency of cfDNA breakpoints comprises calculating the ratios of number of cfDNA fragments starting or the ratio of number of cfDNA fragments ending at a particular position compared to the number of fragments with start or end positions within 50 bp surrounding that particular position.

4. The method of claim 3, wherein cfDNA fragments comprising similar end-position sequences comprise similar motifs.

5. The method of claim 4, wherein the motifs comprise a thymine or an adenine before a start of the cfDNA fragment sequence and two cytosines (A/T|CC) or a cytosine followed by a guanine (A/T|CG) as first two nucleotides of the cfDNA fragment sequence or where these sequences are the inverse complement at the end of a cfDNA fragment.

6. The method of claim 5, wherein the frequency of the motifs is increased or decreased in healthy subjects as compared to a subject having cancer.

7. The method of claim 6, wherein the frequency of A/T|CC motifs is greater than the frequency of A/T|CG motifs.

8. The method of claim 7, wherein the A/T|CG motifs are positioned close to a histone H1 linker or separated between 100 base pairs to 200 base pairs from the histone H1 linker.

9. The method of any one of claims 1 through 8, wherein interior regions of cfDNA fragments are enriched with adenines and thymines.

10. The method of claim 1, further comprising mapping cfDNA fragments to the genome and comparing the cfDNA fragment end sequences to methylated and unmethylated CpG sites of cfDNA from healthy subjects.

11. The method of claim 10, wherein methylated CpGs are enriched at the ends of A/T|CG cfDNA fragment sequences.

12. The method of claim 11, wherein quantitative assessment of enrichment of cfDNA fragment ends at CpGs comprise calculating for each CpG a fraction of cfDNA fragments starting at a CpG dinucleotide and a fraction of cfDNA fragments ending at aCpG dinucleotide over the number of cfDNA fragments with start or end positions within 50 bp around each CpG.

13. The method of any one of claims 1 through 12, wherein methylated CpGs are reduced at the ends of cfDNA fragments that start with a CCG trinucleotide (containing the CpG at location 2 and 3).

14. The method of any one of claims 1 through 13, wherein quantitative assessment of reduction of cfDNA fragment ends at CCGs comprise calculating for each CCG a fraction of cfDNA fragments starting at a CCG trinucleotide over the number of cfDNA fragments in a region around the first cytosine of that CCG trinucleotide (preferably 50 bps upstream and downstream from that particular position). A similar quantitative assessment of reduction of cfDNA fragment ends at the inverse complement of the CCG trinucleotide (CGG) can be made by calculating a fraction of cfDNA fragments ending at a CGG trinucleotide over the number of cfDNA fragments in a region around the last nucleotide of this trinucleotide.

15. The method of any one of claims 1 through 14, wherein cfDNA fragments obtained from an X chromosome among healthy subjects comprise increased cfDNA fragments ending with CG and decreased cfDNA fragments ending with CCG at locations of X chromosome CpG islands in female subjects as compared to male subjects.

16. The method of any one of claims 1 through 15, wherein increased cfDNA fragment-sequence coverage comprise increased methylation across regions of methylated CpG islands.

17. The method of any one of claims 1 through 16, wherein gene expression at transcription start sites (TSS) is inversely related to cfDNA coverage at TSS.

18. The method of any one of claims 1 through 17, wherein decreased CpG-island methylation is related to an increased amount of smaller cfDNA fragments in the regions around the CpG-island.

19. The method of any one of claims 1 through 18, wherein increased gene expression at transcription start sites is related to increased amount of smaller cfDNA fragments in the regions around the transcription start site.

20. The method of any one of claims 1 through 19 further comprising treating the subject with one or more chemotherapeutic agents, radiotherapy, surgery and combinations thereof.

21. A method of diagnosing cancer and treating a subject, comprising;

obtaining a sample from a subject,

assaying for changes in circulating cell-free DNA (cfDNA) fragment sizes as compared to normal and tumor-derived cfDNA controls,

assaying for CpG methylation on cfDNA fragment ends by assessing fragment end representation at CG and CCG sites through low coverage whole genome cfDNA analyses;

diagnosing the subject with cancer; and,

treating the subject with one or more chemotherapeutic agents, radiotherapy, surgery and combinations thereof.

22. The method of claim 21, wherein an increase in cfDNA fragments ending in N|CCG at locations with decreased methylation in the cancer as compared to cfDNA from healthy individuals, is diagnostic of cancer.

23. The method of claim 21 or 22, wherein a decrease in cfDNA fragments ending with N|CCG at locations with increased methylation in the cancer as compared to cfDNA from healthy individuals, is diagnostic of cancer.

24. The method of any one of claims 21 through 23, wherein an increase of cfDNA fragments ending with CG at locations with increased methylation in the cancer as compared to cfDNA from healthy individuals, is diagnostic of cancer.

25. The method of any one of claims 21 through 24, wherein a decrease of cfDNA fragments ending with CG at locations with decreased methylation in the cancer as compared to cfDNA from healthy individuals, is diagnostic of cancer.

26. The method of any one of claims 21 through 25, further comprising incorporation of distribution of fragment end positions at CG and CCG sites in a gradient boosted tree machine learning model.

27. The method of any one of claims 21 through 26, further comprising incorporation of distribution of fragment end positions at CG and CCG sites, and fragment size distributions in a gradient boosted tree machine learning model.

28. A method for diagnosing and treating a subject diagnosed with cancer, comprising:

assaying genomic sequences to identify genome-wide CpG methylation;

analyzing frequency of cfDNA breakpoints at a plurality of positions in the genomic sequences and determining circulating cell-free DNA (cfDNA) fragmentation, wherein, recurrent cfDNA fragment end enrichment at CpG sites correlate with higher genome-wide methylation levels and smaller cfDNA fragments diagnostic of cancer;

diagnosing the subject with cancer if hypomethylation and/or increased gene expression or a decrease in cfDNA fragment size is detected; and,

treating the subject with one or more chemotherapies, radiation, surgery or combinations thereof.

29. The method of claim 28, wherein the genomic sequences are assayed by whole genome sequencing or obtaining whole genome sequences from a database and pooling cfDNA sequences.

30. The method of claim 29, wherein analysis of the frequency of cfDNA breakpoints comprises calculating a ratio of number of cfDNA fragments starting or ending at a particular

position compared to the number of fragments with start or end positions within 50 bp surrounding that particular position.

31. The method of any one of claims 28 through 30, wherein cfDNA fragments comprising similar end-position sequences comprise similar motifs.

32. The method of claim 31, wherein the motifs comprise a thymine or an adenine before a start of the cfDNA fragment sequence and two cytosines (A/T|CC) or a cytosine followed by a guanine (A/T|CG) as first two nucleotides of the cfDNA fragment sequence.

33. The method of any one of claims 28 through 32, wherein the frequency of the motifs is increased in healthy subjects as compared to a subject having cancer.

34. The method of claim 33, wherein the frequency of A/T|CC motifs is greater than the frequency of A/T|CG motifs.

35. The method of claim 33 or 34, wherein the A/T|CG motifs are positioned close to a histone H1 linker or centered between 100 base pairs to 200 base pairs from the histone H1 linker.

36. The method of any one of claims 28 through 35, wherein interior regions of cfDNA fragments are enriched with adenines and thymines.

37. The method of any one of claims 28 through 36, further comprising mapping cfDNA fragments to the genome and comparing the cfDNA fragment end sequences to methylated and unmethylated CpG sites of cfDNA from healthy subjects.

38. The method of claim 37, wherein methylated CpGs are enriched at the ends of A/T|CG cfDNA fragment sequence.

39. The method of any one of claims 28 through 38, wherein quantitative assessment of enrichment of cfDNA fragment ends at CpGs comprise calculating for each CpG a fraction of cfDNA fragments starting or ending at CpG dinucleotide positions over number of cfDNA fragments with start or end positions within 50 bp around each CpG.

40. The method of any one of claims 28 through 39 wherein cfDNA fragments obtained from an X chromosome among healthy subjects comprise increased cfDNA fragments ending with CG and cfDNA fragment ends with CCG are decreased at locations of X chromosome CpG islands in female subjects as compared to male subjects.

41. The method of any one of claims 28 through 40 wherein cfDNA sequence coverage of enriched cfDNA fragment end sequences comprise increased methylation across regions of methylated CpG islands as compared to reduced or lower occurrence cfDNA fragment end sequences.

42. The method of any one of claims 28 through 41, wherein gene expression at transcription start sites (TSS) is inversely related to cfDNA coverage at TSS.