

SINGLE MOLECULE GENOME-WIDE MUTATION AND FRAGMENTATION PROFILES OF CELL-FREE DNA

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH

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

CROSS REFERENCE TO RELATED APPLICATIONS

[0002] This Application claims the benefit of U.S. Provisional Application 63/290,017 filed on December 15, 2021. The entire contents of this application are incorporated herein by reference in its entirety.

FIELD

[0003] Embodiments are directed to methods for determining the frequency of somatic mutations in a subject and in particular the diagnosis and treatment of cancer.

BACKGROUND

[0004] Most of the mortality of human cancer is a consequence of late diagnosis when therapies are less effective¹. Early screening for cancer has demonstrated clinical benefit in multiple cancer types, but implementation of screening approaches remains a challenge². For example, screening for lung cancer using low-dose computed tomography (LDCT) is currently recommended in the United States for adults 50-80 years in age who have smoked at least 20 pack years and currently smoke or have quit within the last 15 years³. Although screening with LDCT has been shown to reduce mortality^{4,5} adherence to this test is low (<6%) among high-risk individuals⁶, in part due to concerns about potential harms from its low specificity, radiation exposure, and unnecessary diagnostic procedures. Liquid biopsies may overcome some of these challenges and provide an attractive approach for non-invasive detection of lung cancer and other malignancies.

SUMMARY

[0005] Provided herein is a non-invasive and an ultrasensitive analysis of single cell-free DNA (cfDNA) molecules to detect the frequency of somatic mutations across the genome.

Patients with cancer were found to have altered mutational profiles associated with chromatin organization compared to healthy individuals.

[0006] Accordingly, in certain aspects, a method of determining the frequency of somatic mutations in a subject, comprises extracting cell-free DNA (cfDNA) from a subject's biological sample; generating genomic libraries from the extracted cfDNA; sequencing individual cfDNA molecules to obtain mutation profiles; determining multiregional differences in mutation profiles; and determining the frequency of somatic mutations in the subject.

[0007] In certain embodiments, the determination of genome-wide mutation and fragmentation profiles comprises identifying mutations in sequences of individual cfDNA molecules and changes in fragment lengths.

[0008] In certain embodiments, the mutation profiles comprise mutation frequency and type of mutation across the subject's genome.

[0009] In certain embodiments, the mutation profiles across the subject's genome are determined using non-overlapping bins ranging in size from at least about one thousand bases to at least about twenty million bases.

[0010] In certain embodiments, the mutation profiles across the subject's genome are determined using non-overlapping bins ranging in size from at least about one thousand bases to at least about ten million bases.

[0011] In certain embodiments, the mutation profiles across the subject's genome are determined using non-overlapping bins ranging in size from at least about one thousand bases to at least about five million bases.

[0012] In certain embodiments, the mutations for each sequenced molecule are determined after removing common germline variants, and unevaluable regions.

[0013] In certain embodiments, the frequency of single molecule somatic mutations and type of mutation across the subject's genome is diagnostic of cancer as compared to the frequency of single molecule somatic mutations and type of mutation across a normal subject's genome.

[0014] In certain aspects, a method of treating cancer in a subject, the method comprises extracting cell-free DNA (cfDNA) from a subject's biological sample; generating genomic libraries from the extracted cfDNA; sequencing individual cfDNA molecules to obtain mutation profiles; determining multiregional differences in mutation profiles and determining the frequency of somatic mutations in the subject; and on the basis thereof administering a cancer treatment to the subject.

[0015] In certain embodiments, the cancer treatment comprises: surgery, adjuvant chemotherapy, neoadjuvant chemotherapy, radiation therapy, hormone therapy, cytotoxic therapy, immunotherapy, adoptive T cell therapy, targeted therapy, and combinations thereof.

[0016] In certain embodiments, the cancer comprises colorectal cancer, lung cancer, breast cancer, gastric cancers, pancreatic cancers, bile duct cancers, brain cancer or ovarian cancer.

[0017] In certain embodiments, the lung cancer is small cell lung cancer (SCLC).

[0018] In certain embodiments, the lung cancer is non-small cell lung cancer (NSCLC).

[0019] In certain embodiments, the subjects with cancer comprise altered mutational profiles associated with chromatin organization as compared to healthy individuals.

[0020] In certain embodiments, the genome-wide mutation and fragmentation profiles comprises identifying mutations in sequences of individual cfDNA molecules and changes in fragment lengths.

[0021] In certain embodiments, the mutation profiles comprise mutation frequency and type of mutation across the subject's genome.

[0022] In certain embodiments, the mutation profiles across the subject's genome are determined using non-overlapping bins ranging in size from at least about one thousand bases to at least about twenty million bases.

[0023] In certain embodiments, the mutation profiles across the subject's genome are determined using non-overlapping bins ranging in size from at least about one thousand bases to at least about ten million bases.

[0024] In certain embodiments, the mutation profiles across the subject's genome are determined using non-overlapping bins ranging in size from at least about one thousand bases to at least about five million bases.

[0025] In certain embodiments, the genome-wide mutations for each sequenced molecule are determined after removing common germline variants, and unevaluable regions.

[0026] In certain embodiments, a method of determining regional frequency of mutations across a genome including sequencing of individual cfDNA molecules isolated from a subject, estimating mutation frequencies and types of mutations across the genome; determining the mutation types and frequencies in genomic regions altered in cancer to mutation profiles and regions mutated in normal cfDNA to determine multiregional differences in mutation profiles; thereby, determining regional frequency of mutations across a genome. In certain embodiments, the estimation of mutation frequencies and types of mutations across the genome comprise using non-overlapping bins ranging in size from thousands to millions of bases. In certain embodiments, tumor specific changes are quantified by one or more assays. In certain embodiments, the one or more assays comprise *in silico* dilution assays and/or downsampling assays. In certain embodiments, each sequenced molecule is scanned for single nucleotide changes after removing common germline variants and/or unevaluable regions. In certain embodiments, the genomic regions are characterized by late replication timing, low gene expression, B compartmentalization, high H3K9me3 abundance, low GC content, or a combination thereof. In certain embodiments, the frequency of putative mutations is defined as the number of variants per million evaluated positions across all the DNA molecules sequenced. In certain embodiments, the method further comprises combining mutational profiles and genome-wide fragmentation profiles. In certain embodiments, the method further comprises executing a machine learning model for determining changes in genome-wide mutational profiles that classifies or excludes the subject as having or at risk of having cancer based on the genome-wide mutational profile identified for the subject.

[0027] In certain embodiments, a method of determining whether a subject is responding to treatment, comprises any one or more of the methods embodied herein. In certain embodiments, the treatment is selected from surgery, adjuvant chemotherapy, neoadjuvant

chemotherapy, radiation therapy, hormone therapy, cytotoxic therapy, immunotherapy, adoptive T cell therapy, targeted therapy, and combinations thereof.

[0028] *Definitions*

[0029] 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.

[0030] 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.”

[0031] 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.

[0032] 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).

[0033] 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 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.

[0034] 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.

[0035] 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.

[0036] 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.

[0037] “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.

[0038] An “effective amount” as used herein, means an amount which provides a therapeutic or prophylactic benefit.

[0039] 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.

[0040] As used herein, the “frequency” of mutations is defined as the number of variants per million evaluated positions across all the DNA molecules sequenced.

[0041] 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. In various embodiments, genomic DNA can be extracted from a cell derived from a blood cell lineage, such as a white blood cell (WBC).

[0042] As used herein, the term “mutational profile” refers to the mutation type and frequency as observed in bins across the genome. Comparison of mutation profiles between genomic regions more commonly altered in cancer and mutation profiles from regions more frequently mutated in normal cfDNA can be used to determine multiregional differences.

[0043] “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.

[0044] 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.

[0045] “Parenteral” administration of an immunogenic composition includes, e.g., subcutaneous (s.c.), intravenous (i.v.), intramuscular (i.m.), or intrasternal injection, or infusion techniques.

[0046] 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.

[0047] 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](https://www.ncbi.nlm.nih.gov).

[0048] 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.

[0049] 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.

[0050] 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.

[0051] 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.

[0052] 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.

[0053] 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.

[0054] 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.

[0055] 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

[0056] **FIG. 1** is a schematic of an overall approach for cancer detection using single molecule cfDNA sequencing. Blood is collected from a population of individuals, some of whom have cancer. cfDNA is extracted from plasma and subject to single molecule sequencing using massively parallel sequencing approaches. Sequence alterations are used to obtain genome-wide mutation profiles and regional differences in cancer and non-cancer mutation frequencies and are identified using machine learning to distinguish individuals with and without cancer.

[0057] **FIGS. 2A-2J** are a series of graphs and plots showing the single molecule mutation analyses of lung cancers from the PCAWG consortium and normal samples. **FIG. 2A:** Number of mutations detected in PCAWG lung cancer samples of smoking individuals when downsampled across a range of sequencing coverage amounts and tumor fractions. **FIG. 2B:** Fraction of PCAWG lung cancer mutations observed in single DNA molecules at the different sequence coverage and tumor fractions indicated. **FIG. 2C:** Frequency of single molecule somatic and background C>A changes in lung cancer and blood derived matched normal samples without quality or germline filters. **FIG. 2D:** Frequency of single molecule somatic and

background C>A changes in lung cancer and blood derived matched normal samples with quality and germline filters including filtering of 8-oxo-dG related sequence changes. **FIG. 2E:** Frequency of single molecule somatic and background C>A changes spanning a 50 Mb region of chromosome 1 in patient DO25320. The C>A frequency was computed in a sliding 2.5 Mb window with a step size of 100 kb. The red and black dashed lines represent the mutation frequencies of the top decile of bins most enriched in C>A changes in lung cancers and matched blood derived normal samples. **FIG. 2F:** Background C>A frequency of the top decile of bins most enriched in C>A changes in lung cancer and matched blood derived normal samples obtained after removal of the known PCAWG somatic mutations. For each sample, the background C>A frequencies are similar between these regions as can be seen with the solid identity line. **FIG. 2G:** Number of molecules with each background C>A change in lung cancer and blood derived normal samples. Most background changes are only observed in a single molecule, even at >30x coverage. **FIG. 2H:** Difference in regional C>A frequencies in normal or tumor samples after subtraction of the C>A frequency in the top decile of bins enriched in normal samples from the top decile of bins enriched in mutations in tumor samples using the GEMINI approach. The difference in regional C>A frequencies preferentially removes background changes, thereby enriching the frequency of the observed somatic mutations. **FIG. 2I:** Association between the regional difference in single molecule C>A frequency and the frequency of high-confidence somatic C>A changes reported in these samples by the PCAWG consortium. **FIG. 2J:** Receiver operating characteristic (ROC) curve for distinguishing lung cancer from normal samples using the GEMINI approach with the testing set downsampled to 1x coverage compared to using overall single molecule C>A frequencies after quality and germline filtering. The ROC for GEMINI approach without filtering 8-oxo-dG related changes results in an AUC of 0.47, highlighting the importance of removing these artifacts for identification of tumor-specific alterations.

[0058] **FIGS. 3A-3B** are a series of plots and graphs demonstrating the genome-wide mutation profiles of tissue and plasma samples were associated with replication timing. **FIG. 3A:** Somatic mutation frequencies PCAWG lung cancers from smoking individuals (n=65) were computed in sliding 2.5 Mb windows with a step size of 100 kb across the genome and represented as the average across individuals. **FIG. 3B:** Association of mutation frequencies across tissue-specific replication timing strata in tissue and cfDNA from patients with NSCLC,

melanoma, BNHL, or no cancer. Replication timing was obtained as the wavelet-smoothed transform of the six fraction profile representing different time points during replication in 1 kb bins from IMR90, NHEK and GM12878 cell lines^{47,48} for analyses of NSCLC, melanoma, and BNHL respectively. The weighted average of the replication timing values was computed in 2.5 Mb bins, followed by grouping of bins into 5 equal bin sets containing bins with the earliest to latest replication timing. In each bin set, we computed the mutation frequency in tissue at different replication strata using the number of somatic mutations reported by the PCAWG Consortium per Mb of genome and compared this to the single molecule mutation frequency in plasma using a Pearson correlation. To control for potential systematic variability in measured genome-wide mutational frequencies, we subtracted from both cancer and non-cancer cfDNA samples the single molecule mutation frequency in each bin set in a separate panel of 20 non-cancer cfDNA samples. Mutation frequencies were then scaled within each sample and mutation type to have a minimum value of zero.

[0059] FIGS. 4A-4I are a series of plots and ROC curves demonstrating the detection of lung cancer using GEMINI and combined GEMINI/DELFI approach. **FIG. 4A:** GEMINI scores in high-risk individuals, age 50-80 with a ≥ 20 pack per year smoking history with or without lung cancer, with the number of individuals indicated at each stage or histology. Importantly, non-cancer individuals with and without benign nodules had similar GEMINI scores, and individuals with cancer had higher GEMINI scores. **FIG. 4B:** GEMINI scores of high-risk individuals without lung cancer as well as individuals without lung cancer as determined by imaging at baseline but who later developed lung cancer. **FIG. 4C:** GEMINI scores in the validation cohort in current or former smokers aged 50-80 with and without cancer. The validation cohort was enriched for early stage disease (stage I=25, stage II=2, stage III=2, stage IV=2, and 1 individual with unknown stage). **FIG. 4D:** ROC curve for detection of lung cancer in high-risk individuals in the LUCAS cohort (n=89 with lung cancer, n=74 without cancer) shows high performance using the GEMINI or GEMINI and DELFI approaches. **FIG. 4E:** ROC curve for detection of lung cancer in subset of high-risk individuals in the LUCAS cohort with at least 40 pack years (n=63 with lung cancer, n=46 without cancer) shows increased performance of GEMINI with higher smoking history. **FIG. 4F:** ROC curve for detection of high-risk individuals from LUCAS cohort who were diagnosed with stage I lung cancer (n=13 with lung cancer, n=74 without cancer). **FIG. 4G:** ROC curve for detection of stage I lung cancer among

individuals in the validation cohort (n=25 with lung cancer, n=14 without cancer). **FIG. 4H:** ROC curve for detection of high-risk individuals from LUCAS cohort and a ≥ 40 pack year smoking history who were diagnosed with stage I lung cancer (n=9 with lung cancer, n=46 without cancer). **FIG. 4I:** ROC curve for detection of stage I lung cancer and a ≥ 40 pack year smoking history among individuals in the validation cohort (n=13 with lung cancer, n=5 without cancer).

[0060] **FIGS. 5A-5F** are a series of graphs and an ROC curve demonstrating the GEMINI approach for noninvasive detection across multiple cancer types. **FIG. 5A:** GEMINI scores in SCLC patients and high-risk individuals without cancer in the LUCAS and the validation cohorts show high performance for detecting cancer (Supplementary Table 4). **FIG. 5B:** Regional differences in single molecule C>A frequency in the LUCAS and validation cohorts demonstrates that the GEMINI approach can be used to identify bins most altered between SCLC and NSCLC. **FIG. 5C:** ROC curves for detection of SCLC (n=13) compared to non-cancer controls (n=88) (orange) as well as for distinguishing SCLC (n=13) from NSCLC (n=99) (purple) in the combined LUCAS and validation cohorts. **FIG. 5D:** Cross-validated regional differences in single molecule mutation frequencies in cfDNA in the liver cancer cohort, median-centered within each mutation type, show a high level of T>C mutations in patients with HCC. P-values were generated using the Wilcoxon rank sum test and were corrected for multiple comparisons using the Benjamini-Hochberg method. The horizontal dashed line indicates a p-value of 0.05. **FIG. 5E:** GEMINI scores in the liver cancer cohort with the number of individuals indicated at each stage demonstrate high sensitivity for detection of liver cancer across all stages. **FIG. 5F:** Principal coordinate analysis of the Euclidean distance matrix reflecting cross-validated pairwise differences in regional mutation frequencies between NSCLC, SCLC, and HCC. The first two principal coordinates are shown with contours indicating kernel density estimations for 0.7 and 0.95 probability for each cancer type. The composition of cancer types in clusters derived from K-means clustering with k=3 is indicated to the right.

[0061] **FIG. 6** is a schematic showing an overview of cohorts analyzed. Each box represents a cohort analyzed and indicates whether the GEMINI approach was evaluated with either cross-validation or validated using a fixed model. Dashed lines indicate analyses of cohort subsets for evaluation of individual tumor types or comparison of cancer subtypes.

[0062] **FIG. 7** is a series of plots showing the genomic mutation profiles in common cancers. Average somatic mutation frequencies computed in sliding 2.5 Mb windows with a step size of 100 kb across chromosome 1 obtained were obtained from an analysis of 2,511 PCAWG samples across 25 common cancer types.

[0063] **FIG. 8** is a schematic of dilution and downsampling experiment. In this example, we consider a tumor sample that contains N somatic mutations at genomic positions 1, 2, ..., N . We begin with 30 non-tumor derived observations and 10 tumor derived mutations (25% tumor purity). During the dilution step, non-tumor observations are spiked in until the desired tumor fraction is achieved. After dilution, fragments are randomly sampled from the set of all fragments to achieve the desired average coverage across genomic positions. The resulting number of observed mutations is counted, and lastly, the proportion of observed mutations that are only observed in a single fragment is computed. In this example, there are 3 observed mutations and one of them is only observed in a single molecule.

[0064] **FIGS. 9A, 9B** is a plot and a graph demonstrating the identification of background changes in single molecule sequencing related to 8-oxo-dG damage. **FIG. 9A:** Ratio of the frequency of each type of single base change in 62 tissue samples from PCAWG (31 lung cancer and 31 blood derived matched normal samples) when prior to mutation purines guanine or adenine (pu) are on read 1 (R1) or pyrimidines cytosine or thymine (py) are on read 2 (R2) to when the pyrimidine is on read 1 and the purine is on read 2 for both background changes and known germline variants. Background changes reflect sequence changes identified through single molecule analyses that were not reported as somatic variants by PCAWG. Here, germline variants reported by PCAWG were also removed from the background variants to enrich for likely artifactual changes. **FIG. 9B:** Ratio of known somatic mutations to background changes identified through single molecule analyses before removing likely 8-oxo-dG related sequence changes ($R1_{pu \text{ or } py}$, $R2_{pu \text{ or } py}$), and after filtering these changes where only bases with cytosine on R1 and guanine on read 2 are considered ($R1_{py}$, $R2_{pu}$).

[0065] **FIGS. 10A, 10B** are plots demonstrating the analyses of single molecule sequence changes in PCAWG lung cancer and normal samples. **FIG. 10A:** Single molecule mutation frequencies in PCAWG lung cancers ($n=31$) and blood derived matched normal samples ($n=31$). P-values were corrected for multiple comparisons using the Benjamini-

Hochberg method. The horizontal dashed line indicates a p-value of 0.05. **FIG. 10B:** Cross-validated regional differences in single molecule mutation frequencies in PCAWG lung cancers (n=31) and blood derived matched normal samples (n=31), median-centered within each mutation type. P-values were generated using the Wilcoxon rank sum test and were corrected for multiple comparisons using the Benjamini-Hochberg method. The horizontal dashed line indicates a p-value of 0.05.

[0066] **FIG. 11** is a graph demonstrating the analysis of somatic and background changes across mutation types in PCAWG lung cancers. Ratio of somatic to background changes identified through single molecule analyses after removal of potential 8-oxo-dG related artifacts for each mutation type analyzed. Somatic changes reflect sequence changes identified through single molecule analyses that were also reported as somatic mutations by PCAWG, whereas background changes were identified through single molecule analyses but not reported as somatic mutations by PCAWG. Overall, C>G>A:T changes represented the highest fraction of somatic changes.

[0067] **FIG. 12** is a series of plots demonstrating the analysis of single molecule sequence changes across sequencing lanes in PCAWG lung cancer and normal samples. Single molecule mutation frequencies in PCAWG lung cancer and blood derived normal samples across sequencing lanes. For each sample, sequencing reads were split into separate Binary Alignment Map (BAM) files based on their associated read group, which indicates the sequencing reads from one lane of an NGS experiment. The resulting BAM files contained a median of 464 million reads (range: 6-738 million). Approximately 1 million reads were randomly sampled 5 times with replacement from each sequencing lane (a maximum of 6 lanes is shown per sample). Single molecule mutation frequencies varied widely within an individual sample depending on the analyzed lane and the type of sequence alteration.

[0068] **FIG. 13** is a series of plots demonstrating the genome-wide somatic single molecule C>A mutation profiles in lung cancers. Single molecule C>A somatic mutation frequencies computed in sliding 2.5 Mb windows with a step size of 100 kb across the autosomes obtained from an aggregated analysis of the 31 PCAWG lung cancer samples showed widespread differences in mutation frequencies depending on genomic location.

[0069] **FIG. 14** is a series of plots demonstrating somatic single molecule C>A mutation profiles across chromosome 4 in PCAWG lung cancers. Single molecule C>A somatic mutation frequencies computed in a sliding 2.5 Mb window with a step size of 100 kb across chromosome 4 from PCAWG lung cancer samples revealed similar mutation profiles among different lung cancers.

[0070] **FIG. 15** is a schematic of a GEMINI regional mutation frequency analysis. The genome is divided into 1,144 non-overlapping 2.5Mb bins (20 bins are depicted here) and the single molecule mutation frequency is computed in each bin as the number of sequence changes per million evaluable bases, defined as the number of positions in fragments in which each sequence change could be detected after quality and germline filtering. Samples in the training set are used to identify the bins that are most differentially mutated between cancer and non-cancer samples. In the training set, sequence data from all cancer samples and all non-cancer samples are combined, and the cancer and non-cancer single molecule mutation frequencies are computed in each bin. Next, the difference in single molecule mutation frequency is computed between cancer and non-cancer samples in each bin, and the 10% of bins most mutated in cancer samples relative to non-cancer samples, as well as the 10% of bins most mutated in non-cancer samples relative to cancer samples, are identified (indicated by triangles and circles respectively). In the testing set, the difference in single molecule mutation frequency is computed between these two sets of bins in a new sample not included in the training set, generating a regional difference in mutation frequency that can be used to classify the sample into being derived from a healthy individual or an individual with cancer. By taking the difference in single molecule mutation frequency between two sets of regions in the genome within an individual sample, this approach controls for the overall number of sequence changes in that sample that may result from technical variability in sequencing runs.

[0071] **FIG. 16** is a graph demonstrating the effect of matched WBC filtering in PCAWG lung cancers on enrichment of somatic alterations by single molecule sequencing. Single molecule C>A frequency in PCAWG lung cancers (n=31) after removal of any sequence changes identified in matched blood derived normal samples at >30x coverage. The analysis revealed that subtraction of mutations observed in the matched normal sample was not effective

in removing background changes because such alterations typically were observed once and were not present in both tumor and matched non-cancer samples.

[0072] **FIGS. 17A-17C** are a series of plots demonstrating the association of single molecule genome-wide mutation profiles of tissue and plasma samples with genomic features. The figures show the genome-wide mutation frequencies across strata of tissue-specific gene expression, A/B compartmentalization, and H3K9me3 abundance, respectively, in tissue and cfDNA from patients with NSCLC, melanoma, BNHL, or without cancer. The weighted average of each feature value was computed in 2.5 Mb bins, followed by grouping of bins into 5 equal bin sets ordered by feature value. In each bin set, we computed the mutation frequency in tissue at different strata using the number of somatic mutations reported by the PCAWG Consortium per Mb of genome and compared this to the single molecule mutation frequency in plasma using a Pearson correlation. To account for difference in the overall frequency of each mutation type in each bin in cfDNA, the single molecule mutation frequency in each bin set in a panel of non-cancer samples (n=20) was subtracted from the single molecule mutation frequency in each bin set in cancer and non-cancer cfDNA samples and the resulting values were scaled to have a minimum value of zero for each mutation type and sample type. **FIG. 17A:** Gene expression was computed as the sum of the transcripts per million (TPM) overlapping each 2.5 Mb bin weighted by the length of the transcript averaged across TCGA NSCLC, melanoma, and BNHL samples. **FIG. 17B:** A/B compartmentalization, largely representing open and closed regions of the genome, respectively, was measured as the first eigenvector of the correlation matrix of average methylation beta values in 100 kb bins across TCGA NSCLC samples for NSCLC analyses and was averaged across 12 TCGA cancer types for melanoma analyses. The first eigenvector for the genome contact matrix from Hi-C analyses of lymphoblastoid cells (GM12878 cell line) was used for BNHL analyses³³. **FIG. 17C:** The abundance of H3K9me3, a known marker of heterochromatin, was obtained from ChIP-seq of A549 cells (three pooled replicates), GM23248, and Karpas 422 cells (two pooled replicates) for NSCLC, melanoma, and BNHL analyses respectively as the fold change of coverage in enriched samples compared to control samples⁴⁸.

[0073] **FIG. 18** is a plot demonstrating the regional differences in single molecule mutation frequencies in the high-risk LUCAS cohort. Cross-validated regional differences in single molecule mutation frequencies in cfDNA in individuals with lung cancer (n=89) and

individuals without cancer (n=74), median-centered within each mutation type. Regional C>A mutation frequencies were preferentially altered between lung cancer and non-cancer samples, but not when randomly permuting class labels (p=0.36, Wilcoxon rank sum test). P-values were generated using the Wilcoxon rank sum test and were corrected for multiple comparisons using the Benjamini-Hochberg method. The horizontal dashed line indicates a p-value of 0.05.

[0074] **FIG. 19** is a plot showing the analyses of C>A sequence changes by flow cell and sequencing lane in non-cancer individuals. Single molecule C>A frequencies and regional differences in single molecule C>A frequencies across flow cells and sequencing lanes for all non-cancer individuals from the LUCAS cohort (n=158). Although sequencing background mutation rates differed by lane resulting in multiple samples within a sequencing lane having similar single molecule C>A frequencies, explaining 99% of their variance (p<0.0001, F-test), this association was eliminated using regional differences in single molecule C>A frequencies obtained with the GEMINI approach (p=0.17, F-test).

[0075] **FIGS. 20A-20K** is a series of plots and schematics showing the genome-wide fixed bins utilized for analysis of single molecule mutation frequencies and detection of lung cancer in cfDNA. **FIG. 20A:** Percent similarity of bins identified as being enriched for mutations in lung cancer and non-cancer samples in each training fold compared to the sets of bins utilized in the fixed model that were identified from analyses of all samples. A high similarity across training folds indicated that bin selection was not driven by individual samples. **FIG. 20B:** Chromosomal location of bins enriched in mutations in cfDNA of patients with lung cancer and bins enriched in mutations in cfDNA of individuals without cancer. **FIG. 20C:** Compared to samples from individuals without cancer, samples from those with lung cancer had more C>A changes per genomic bin across samples in bins enriched in lung cancer and fewer of these changes in bins enriched in non-cancer. **FIGS. 20D-20E:** The average number of evaluable bases and copy number per genomic bin was similar in non-cancer individuals and individuals with lung cancer in bins enriched in lung cancer and bins enriched in non-cancer. Copy number was estimated using ichorCNA. **FIGS. 20F-20K:** Bins in the fixed model were associated with replication timing, gene expression, A/B compartmentalization, and H3K9me3 abundance, GC content, but not sequence mappability. **FIG. 20F:** Replication timing was obtained as the wavelet-smoothed transform of the six fraction profile representing different time points

during replication in 1 kb bins from IMR90 cells^{47,49} and then computing the weighted average in each 2.5 Mb bin with higher values indicating earlier replication timing. **FIG. 20G:** Gene expression was computed as the sum of the transcripts per million (TPM) overlapping each 2.5 Mb bin weighted by the length of the transcript averaged across TCGA NSCLC samples and log transformed as $\log_{10}(\text{TPM})$. **FIG. 20H:** A/B compartmentalization, largely representing open and closed regions of the genome, respectively, was measured as the first eigenvector of the correlation matrix of average methylation beta values in 100 kb bins across TCGA lung cancer samples³³. **FIG. 20I:** The abundance of H3K9me3, a known marker of heterochromatin, was obtained from ChIP-seq of A549 cells⁴⁸ and displayed as the fold change of coverage in enriched samples compared to control samples from three pooled replicates. **FIG. 20J:** GC content in each genomic bin was obtained from the hg19 reference genome. Bins enriched in lung cancer tend to be AT rich (GC poor) compared to bins enriched in non-cancer, which may be explained through our previous results that later replicating regions which are enriched in mutations in lung cancer have lower GC content (Spearman's $\rho = 0.83$, $p < 0.0001$). **FIG. 20K:** Mappability, reflecting how uniquely 100-mer sequences align to a region of the genome, was computed as the weighted average in 2.5Mb bins.

[0076] FIGS 21A-21F are a series of plots showing an analyses of doublet base substitutions in tissue and plasma samples of lung cancer patients. **FIG. 21A:** Number of somatic doublet base substitutions identified by the PCAWG Consortium in lung cancer tissue samples from smoking individuals ($n=65$) revealed a high number of CC>AA changes compared to other doublet mutations. Solid horizontal lines indicate the median number of each mutation type across individuals. **FIG. 21B:** Ratio of single molecule CC>AA frequencies when CC or CC>AA is in read 1 and GG or GG>TT is in read 2 (R_{1CC} , R_{2GG}) relative to when GG or GG>TT is in read 1 and CC or CC>AA is in read 2 (R_{1GG} , R_{2CC}) aggregated across samples in the high-risk LUCAS cohort. Background CC>AA changes represent those alterations that were only observed in single cfDNA fragments in individuals without cancer, whereas likely somatic changes represent those alterations that are private to an individual sample from a patient with lung cancer and are observed in multiple cfDNA fragments. Within the high-risk LUCAS cohort, there were 67 private CC>AA changes that were observed in two or more fragments across 89 individuals with lung cancer, and only one such changes observed across 74 individuals without cancer, indicating that most of these alterations were likely somatic in origin. Bars represent 95%

bootstrap confidence intervals for the ratios. Background CC>AA changes were more often detected as R1_{CC}, R2_{GG}, but no imbalance in likely somatic CC>AA changes was observed, indicating an enrichment of likely artifactual background CC>AA changes detected as R1_{CC}, R2_{GG}. **FIG. 21C:** Sequence context surrounding CC>AA changes ($\pm$ 5bp) in the high-risk LUCAS cohort, where the number of mutations is indicated for each group, and the total height of the letters at each position indicates the information content of the position measured in bits. **FIG. 21D:** Single molecule CC>AA frequencies were elevated in individuals with lung cancer compared to non-cancer individuals with a larger separation observed after filtering CC>AA changes detected as R1_{GG}, R2_{CC}. **FIGS. 21E-21F:** Single molecule CC>AA frequencies were positively correlated with regional differences in single molecule C>A frequencies in cfDNA (**FIG. 21E**) and lung tumors (**FIG. 21F**) after filtering CC>AA changes detected as R1_{CC}, R2_{GG}.

[0077] **FIGS. 22A-22F** are a series of plots demonstrating the effect of clinical characteristics on GEMINI scores in non-cancer individuals in the LUCAS cohort. GEMINI scores are shown for **FIG. 22A**, males (n=87) and females (n=71), **FIG. 22B**, individuals with (n=43) or without (n=115) autoimmune disease, **FIG. 22C**, individuals with (n=28) or without (n=130) COPD, **FIG. 22D**, individuals of different ages, and **FIGS. 22E-22F**, compared to CRP (mg/L), and IL-6 levels (pg/mL).

[0078] **FIGS. 23A, 23B** are a series of plots demonstrating that GEMINI scores reflect tumor DNA content in cfDNA. **FIG. 23A:** GEMINI scores in the high-risk LUCAS cohort in individuals without cancer and individuals with lung cancer at different levels of ctDNA. A score >0.55 reflects a positive test for detection of lung cancer at 80% specificity. b, GEMINI scores in the liver cancer cohort in individuals with cirrhosis and individuals with liver cancer that have <3% or $\geq$ 3% ctDNA. A score >0.86 reflects a positive test for detection of liver cancer at 80% specificity. The percentage of ctDNA in each sample was estimated using ichorCNA.

[0079] **FIGS. 24A, 24B** are a series of ROC curves demonstrating the performance of GEMINI or the combined GEMINI / DELFI approach for detection of lung cancer. **FIG. 24A:** ROC curves for detection of lung cancer in the high-risk LUCAS cohort using GEMINI or the combined GEMINI / DELFI approach in patients with stages II-IV disease and in the subset of these patients that smoked $\geq$ 40 pack years. **FIG. 24B:** ROC curves for detection of lung cancer in the high-risk LUCAS cohort using GEMINI or the combined GEMINI / DELFI approach in

patients with adenocarcinoma, squamous cell carcinoma, or small cell lung cancer and in the subset of these patients that smoked ≥ 40 pack years. Performance for Stage I disease is shown in FIGS. 4F, 4H.

[0080] **FIG. 25** is a graph demonstrating GEMINI and DELFI scores as well as their combined performance for detecting cancer in the LUCAS cohort. GEMINI and DELFI scores are shown for each patient in the high-risk LUCAS cohort (n=163). Vertical and horizontal dashed lines indicate the threshold for a positive GEMINI and DELFI test respectively at 80% specificity, while filled circles indicate a positive test by the combined approach at the same specificity. Several individuals with cancer are detected by one approach but not the other, and the combined score detected more individuals with lung cancer compared to either approach in isolation.

[0081] **FIG 26** is a graph demonstrating the GEMINI / DELFI score and clinical outcome in lung cancer patients. Patients with lung cancer in the high-risk LUCAS cohort (n=89) were stratified in two groups based on the median GEMINI / DELFI score among lung cancer patients of 0.84. Patients with a GEMINI / DELFI score ≥ 0.84 (yellow) had a significantly worse overall survival compared to patients with a GEMINI / DELFI score < 0.84 (blue) (p=0.004, Log-rank test).

[0082] **FIGS. 27A-27D** are a series of graphs and plots showing a comparison of cfDNA characteristics across non-cancer patients in LUCAS, DECAMP, and AHN cohorts. **FIG. 27A:** Average genome-wide coverage in non-cancer samples across cohorts. The horizontal dashed lines represent the median coverage of samples in each cohort. **FIG. 27B:** Regional differences in single molecule C>A frequencies are similar between cohorts (p=0.17, Kruskal-Wallis test). Solid horizontal lines represent the median in each group. **FIG. 27C:** For each non-cancer sample, the ratio of short (100-150bp) to long (151-220bp) fragments were computed in 473 non-overlapping 5Mb bins and mean-centered. Median fragmentation profiles represent the median of these values across samples in each bin and are highly correlated between cohorts (Pearson correlation coefficient > 0.97 for each pairwise comparison). **FIG. 27D:** Chromosomal arm-level Z-scores in non-cancer samples are similar between cohorts (p>0.05 for each chromosomal arm, Kruskal-Wallis test with Bonferroni correction).

[0083] **FIGS. 28A-28C** are a series of graphs demonstrating GEMINI scores and smoking exposure in lung cancer patients. **FIG. 28A:** Single molecule C>A frequencies were similar in never smokers with lung cancer (n=3) or without lung cancer (n=34) in the LUCAS cohort. In current or former smokers in the high-risk group, with a ≥ 20 pack year smoking history and age 50-80, the single molecule C>A frequencies were slightly higher in individuals with lung cancer (n=89) compared to individuals without lung cancer (n=74). **FIG. 28B:** GEMINI scores were similar in never smokers with lung cancer (n=3) or without lung cancer (n=34). In the high-risk group, GEMINI scores were higher in individuals with lung cancer (n=89) compared to those without lung cancer (n=74). Similarly, for individuals with a ≥ 40 pack year smoking history and age 50-80, the GEMINI scores were higher in those with lung cancer (n=63) compared to those without lung cancer (n=46). **FIG. 28C:** GEMINI scores were higher in individuals with lung cancer in the validation cohort in current/former smokers age 50-80 with (n=32) and without lung cancer (n=14) and in the subset with a ≥ 40 pack year smoking history with (n=18) and without lung cancer (n=5).

[0084] **FIG. 29** is a graph showing a principal coordinate analysis in patients with cancer after excluding the most frequent mutation types. The regional difference in single molecule mutation frequency was computed between NSCLC, SCLC, and HCC using a leave-one-out procedure for C>G, C>T, T>A and T>G mutations, yielding 12 feature values. A Euclidean distance matrix reflecting pairwise differences between samples was generated from these 12 feature values. A principal coordinate analysis of the Euclidean distance matrix revealed a reduced separation of samples by cancer type compared to when C>A and T>C mutations were also analyzed (FIG. 5F).

[0085] **FIG. 30** is a series of graphs showing GEMINI scores and MAF levels during therapy. Individuals with a smoking history as well as availability of targeted deep sequencing¹¹ and low coverage whole-genome sequencing data¹³ were analyzed before and during treatment with tyrosine kinase inhibitors (arrows indicate initiation of treatment). GEMINI scores were associated with the maximum mutant allele fraction at each timepoint (Spearman's $\rho = 0.50$, $p=0.03$).

DETAILED DESCRIPTION

[0086] Somatic mutations are a hallmark of tumorigenesis and may be useful for non-invasive diagnosis of cancer. However, the detection of somatic alterations in the circulation has been challenging due to the limited number of tumor derived molecules in cell-free DNA (cfDNA). An ultrasensitive analysis of single cfDNA molecules was developed herein, to detect the frequency of somatic mutations across the genome and found that patients with cancer had altered mutational profiles associated with chromatin organization compared to healthy individuals. Combining genome-wide cfDNA mutational profiles and fragmentation features followed by CT imaging detected 95% of patients with cancer across stages and subtypes, including 95% of stage I and II patients, with a 90% combined specificity. The model was independently validated in a separate screening cohort of high-risk individuals with early stage lung cancer. Genome-wide mutational profiles distinguished individuals with small cell lung cancer from those with non-small cell lung cancer and could identify lung cancers earlier than standard approaches. This approach lays the groundwork for non-invasive cancer detection using a combination of genome-wide mutation and fragmentation features of cfDNA that may facilitate cancer screening.

[0087] GEMINI

[0088] Sequence alterations are abundant in cancer genomes but the proportion of fragments in cell-free DNA (cfDNA) that harbor tumor-specific (somatic) mutations is often low^{7,8}, making it difficult to detect *bona fide* variants amongst the background noise due to sequence changes introduced in library construction, gene selection, PCR amplification and sequencing. Extensive efforts have been made to detect mutations that are present at low frequencies in cfDNA. However, these methods typically rely on deep sequencing and have been restricted to examining specific genes comprising a small subset of the genome⁹⁻¹¹. Due to the low number of genome equivalents derived from cancer cells in cfDNA, such approaches have limited efficacy for detecting the presence of cancer especially in early stage disease¹²⁻¹⁴. Additionally, sequence alterations in cfDNA may arise from white blood cells (WBCs), confounding the use of sequence mutations to detect patients with cancer^{7,15,16}.

[0089] The method disclosed herein and termed genome-wide mutational incidence for noninvasive detection of cancer (GEMINI), identified a much larger number of tumor-derived

alterations in cfDNA for cancer detection (FIG. 1). The method is based on sequencing individual cfDNA molecules to estimate the mutation frequency and type of alteration across the genome using non-overlapping bins ranging in size from thousands to millions of bases. For each individual, the mutational profile in genomic regions more commonly altered in cancer is compared to the profile from regions more frequently mutated in normal cfDNA to determine multiregional differences in mutation profiles. In this way, the GEMINI approach enriches for likely somatic mutations while taking into account individual variability in overall mutation number.

[0090] Accordingly, in certain embodiments, a method of determining the frequency of somatic mutations in a subject comprises extracting cell-free DNA (cfDNA) from a subject's biological sample; generating genomic libraries from the extracted cfDNA; sequencing individual cfDNA molecules to obtain mutation profiles; determining multiregional differences in mutation profiles; and, determining the frequency of somatic mutations in the subject.

[0091] The generation of genome-wide mutation profiles included identifying mutations in sequences of individual cfDNA molecules. The mutation profiles across the subject's genome are determined using non-overlapping bins ranging in size from at least about one hundred bases to at least about twenty million bases. In certain embodiments, the mutation profiles across the subject's genome are determined using non-overlapping bins ranging in size from at least about 500 bases to at least about fifteen million bases. In certain embodiments, the mutation profiles across the subject's genome are determined using non-overlapping bins ranging in size from at least about 750 bases to at least about ten million bases. In certain embodiments, the mutation profiles across the subject's genome are determined using non-overlapping bins ranging in size from at least about 900 bases to at least about ten million bases. In certain embodiments, the mutation profiles across the subject's genome determined using non-overlapping bins ranging in size from at least about one thousand bases to at least about five million bases.

[0092] In certain embodiments, the frequency of single molecule somatic mutations and type of mutation across the subject's genome is diagnostic of cancer as compared to the frequency of single molecule somatic mutations and type of mutation across a normal subject's genome.

[0093] In certain embodiments, the frequency of a somatic mutation at various loci is indicative of cancer. In certain embodiments, the type of mutation is indicative of cancer.

[0094] *cfDNA Fragmentation Profiles:* 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.

[0095] 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).

[0096] 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.

[0097] 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).

[0098] Methods of Treatment

[0099] 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 mutation profiles; determining multiregional differences in mutation profiles and determining frequency of somatic mutations in the subject; and administering a cancer treatment to the subject.

[00100] In certain embodiments, a subject is diagnosed as having cancer, *e.g.* early stage cancer. In certain embodiments, the type of cancer is identified, and the cancer is treated by various therapeutics, including therapeutics specific for the type of cancer. In certain embodiments, the cancer comprises colorectal cancer, lung cancer, breast cancer, gastric cancers, pancreatic cancers, bile duct cancers, brain cancer or ovarian cancer. In certain embodiments, the lung cancer is small cell lung cancer (SCLC). In certain embodiments, the lung cancer is non-small cell lung cancer (NSCLC).

[00101] The cancer treatment can be surgery, adjuvant chemotherapy, neoadjuvant chemotherapy, radiation therapy, hormone therapy, cytotoxic therapy, immunotherapy, adoptive T cell therapy, targeted therapy, or any combinations thereof. The method also can include administering to the mammal a cancer treatment (e.g., surgery, adjuvant chemotherapy, neoadjuvant chemotherapy, radiation therapy, hormone therapy, cytotoxic therapy, immunotherapy, adoptive T cell therapy, targeted therapy, or any combinations thereof). The mammal can be monitored for the presence of cancer after administration of the cancer treatment.

[00102] 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, gemcitabine, navelbine, farnesyl-protein transferase inhibitors, transplatin, 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.

[00103] Yet further combination chemotherapies include, for example, alkylating agents such as thiotepa and cyclophosphamide; alkyl sulfonates such as busulfan, improsulfan and piposulfan; aziridines such as benzodopa, carboquone, meturedopa, and uredopa; ethylenimines and methylamelamines including altretamine, triethylenemelamine, trietylenephosphoramide, triethylenethiophosphoramide 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 ranimustine; 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, carabycin, 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, rodornbicin, streptonigrin, streptozocin, tubercidin, ubenimex, zinostatin, zombicin; 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, mepitiothane, testolactone; anti-adrenals such as mritotane, trilostane; folic acid replenisher such as frolinic acid; aceglatone; aldophosphamide glycoside; arninolevulinic acid; eniluracil; arnsacrine; bestrabucil; bisantrene; edatraxate; defofarnine; 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 RPS 2000; difluoromethylornithine (DMFO); retinoids such as retinoic acid; capecitabine; carboplatin, procarbazine, plicomycin, gemcitabine, navelbine, farnesyl-protein transferase inhibitors, transplatin; and pharmaceutically acceptable salts, acids or derivatives of any of the above.

[00104] 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.

[00105] 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 CDS effector memory T cells are characterized by having the expression of CD44^{high} CD62L^{low}.

[00106] 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. Mutant protein antigens are likely to be much more specific to cancer cells because normal cells shouldn't contain these proteins. Normal cells will display the normal protein antigen on their MHC molecules, whereas cancer cells will display the mutant version. Some viral proteins are implicated in forming cancer, and some viral antigens are also cancer antigens. Cancer-testis antigens are antigens expressed primarily in the germ cells of the testes, but also in fetal ovaries and the trophoblast. Some cancer cells aberrantly express these proteins and therefore present these antigens, allowing attack by T-cells specific to these antigens. Exemplary antigens of this type are CTAG1 B and MAGEA1 as well as Rindopepimut, a 14-mer intradermal injectable peptide vaccine targeted against epidermal growth factor receptor vlll (EGFRvlll; deletion of exons 2–7) variant. Rindopepimut is particularly suitable for treating glioblastoma when used in combination with an inhibitor of the CD95/CD95L signaling system as described herein. Also, proteins that are normally produced in very low quantities, but whose production is dramatically increased in cancer cells, may trigger an immune response. An example of such a protein is the enzyme tyrosinase, which is required for melanin production. Normally tyrosinase is produced in minute quantities but its levels are very much elevated in melanoma cells. Oncofetal antigens are another important class of cancer antigens. Examples are alphafetoprotein (AFP) and

carcinoembryonic antigen (CEA). These proteins are normally produced in the early stages of embryonic development and disappear by the time the immune system is fully developed. Thus, self-tolerance does not develop against these antigens. Abnormal proteins are also produced by cells infected with oncoviruses, e.g. EBV and HPV. Cells infected by these viruses contain latent viral DNA which is transcribed, and the resulting protein produces an immune response. A cancer vaccine may include a peptide cancer vaccine, which in some embodiments is a personalized peptide vaccine. In some embodiments, the peptide cancer vaccine is a multivalent long peptide vaccine, a multi-peptide vaccine, a peptide cocktail vaccine, a hybrid peptide vaccine, or a peptide-pulsed dendritic cell vaccine.

[00107] 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, carcinoembryonic antigen, Prostate-specific membrane antigen, Mucin-1, CD30, CD33, or CD40.

[00108] 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-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™ (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 Fe); Abatacept (Extracellular domain of CTLA-4 fused to Fe); 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- α 4 integrin); Vedolizumab (anti- α 4 β 7 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).

[00109] Systems

[00110] In some examples, the present disclosure provides systems, methods, or kits that can include data analysis realized in measurement devices (e.g., laboratory instruments, such as a sequencing machine), software code that executes on computing hardware. The software can be stored in memory and execute on one or more hardware processors. The software can be organized into routines or packages that can communicate with each other. A module can comprise one or more devices/computers, and potentially one or more software routines/packages that execute on the one or more devices/computers. For example, an analysis application or system can include at least a data receiving module, a data pre-processing module, a data analysis module (which can operate on one or more types of genomic data), a data interpretation module, or a data visualization module.

[00111] The data receiving module can connect laboratory hardware or instrumentation with computer systems that process laboratory data. The data pre-processing module can perform operations on the data in preparation for analysis. Examples of operations that can be applied to the data in the pre-processing module include affine transformations, denoising operations, data cleaning, reformatting, or subsampling. The data analysis module, which can be specialized for analyzing genomic data from one or more genomic materials, can, for example, take assembled genomic sequences and perform probabilistic and statistical analysis to identify abnormal patterns related to a disease, pathology, state, risk, condition, or phenotype. The data interpretation module can use analysis methods, for example, drawn from statistics, mathematics, or biology, to support understanding of the relation between the identified abnormal patterns and health conditions, functional states, prognoses, or risks. The data analysis module and/or the data interpretation module can include one or more machine learning models, which can be implemented in hardware, e.g., which executes software that embodies a machine learning model. The data visualization module can use methods of mathematical modeling, computer graphics, or rendering to create visual representations of data that can facilitate the understanding or interpretation of results. The present disclosure provides computer systems that are programmed to implement methods of the disclosure.

[00112] In some embodiments, the methods disclosed herein can include computational analysis on nucleic acid sequencing data of samples from an individual or from a plurality of

individuals. An analysis can identify a variant inferred from sequence data to identify sequence variants based on probabilistic modeling, statistical modeling, mechanistic modeling, network modeling, or statistical inferences. Non-limiting examples of analysis methods include principal component analysis, autoencoders, singular value decomposition, Fourier bases, wavelets, discriminant analysis, regression, support vector machines, tree-based methods, networks, matrix factorization, and clustering. Non-limiting examples of variants include a germline variation or a somatic mutation. In some examples, a variant can refer to an already-known variant. The already-known variant can be scientifically confirmed or reported in literature. In some examples, a variant can refer to a putative variant associated with a biological change. A biological change can be known or unknown. In some examples, a putative variant can be reported in literature, but not yet biologically confirmed. Alternatively, a putative variant is never reported in literature, but can be inferred based on a computational analysis disclosed herein. In some examples, germline variants can refer to nucleic acids that induce natural or normal variations.

[00113] In certain embodiments, the computer system includes a central processing unit (CPU, also “processor” and “computer processor” herein), which can be a single core or multi core processor, or a plurality of processors for parallel processing; memory (e.g., cache, random-access memory, read-only memory, flash memory, or other memory); electronic storage unit (e.g., hard disk), communication interface (e.g., network adapter) for communicating with one or more other systems; and peripheral devices, such as adapters for cache, other memory, data storage and/or electronic display. The memory, storage unit, interface and peripheral devices may be in communication with the CPU through a communication bus (solid lines), such as a motherboard. The storage unit can be a data storage unit (or data repository) for storing data. One or more analyte feature inputs can be entered from the one or more measurement devices. Example analytes and measurement devices are described herein.

[00114] The computer system can be operatively coupled to a computer network (“network”) with the aid of the communication interface. The network can be the Internet, an internet and/or extranet, or an intranet and/or extranet that is in communication with the Internet. The network in some cases is a telecommunication and/or data network. The network can include one or more computer servers, which can enable distributed computing, such as cloud computing over the network (“the cloud”) to perform various aspects of analysis, calculation, and generation

of the present disclosure, such as, for example, activation of a valve or pump to transfer a reagent or sample from one chamber to another or application of heat to a sample (e.g., during an amplification reaction), other aspects of processing and/or assaying a sample, performing sequencing analysis, measuring sets of values representative of classes of molecules, identifying sets of features and feature vectors from assay data, processing feature vectors using a machine learning model to obtain output classifications, and training a machine learning model (e.g., iteratively searching for optimal values of parameters of the machine learning model). Such cloud computing may be provided by cloud computing platforms such as, for example, Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform, and IBM cloud. The network, in some cases with the aid of the computer system, can implement a peer-to-peer network, which may enable devices coupled to the computer system to behave as a client or a server.

[00115] The CPU can execute a sequence of machine-readable instructions, which can be embodied in a program or software. The instructions can be stored in a memory location, such as the memory. The instructions can be directed to the CPU, which can subsequently program or otherwise configure the CPU to implement methods of the present disclosure. The CPU can be part of a circuit, such as an integrated circuit. One or more other components of the system can be included in the circuit. In some cases, the circuit is an application specific integrated circuit (ASIC).

[00116] The storage unit can store files, such as drivers, libraries and saved programs. The storage unit can store user data, e.g., user preferences and user programs. The computer system in some cases can include one or more additional data storage units that are external to the computer system, such as located on a remote server that is in communication with the computer system through an intranet or the Internet.

[00117] The computer system can communicate with one or more remote computer systems through the network. For instance, the computer system can communicate with a remote computer system of a user. Examples of remote computer systems include personal computers (e.g., portable PC), slate or tablet PC's (e.g., Apple® iPad, Samsung® Galaxy Tab), telephones, Smart phones (e.g., Apple® iPhone, Android-enabled device, Blackberry®), or personal digital assistants. The user can access the computer system via the network.

[00118] Methods as described herein can be implemented by way of machine (e.g., computer processor) executable code stored on an electronic storage location of the computer system such as, for example, on the memory or electronic storage unit. The machine executable or machine-readable code can be provided in the form of software. During use, the code can be executed by the CPU. In some cases, the code can be retrieved from the storage unit and stored on the memory for ready access by the CPU. In some situations, the electronic storage unit can be precluded, and machine-executable instructions are stored on memory.

[00119] The code can be pre-compiled and configured for use with a machine having a processor adapted to execute the code or can be compiled during runtime. The code can be supplied in a programming language that can be selected to enable the code to execute in a pre-compiled or as compiled fashion.

[00120] Aspects of the systems and methods provided herein, such as the computer system, can be embodied in programming. Various aspects of the technology can be thought of as “products” or “articles of manufacture” typically in the form of machine (or processor) executable code and/or associated data that is carried on or embodied in a type of machine readable medium. Machine-executable code can be stored on an electronic storage unit, such as memory (e.g., read-only memory, random-access memory, flash memory) or a hard disk. “Storage” type media can include any or all of the tangible memory of the computers, processors or the like, or associated modules thereof, such as various semiconductor memories, tape drives, disk drives and the like, which may provide non-transitory storage at any time for the software programming. All or portions of the software may at times be communicated through the Internet or various other telecommunication networks. Such communications, for example, may enable loading of the software from one computer or processor into another, for example, from a management server or host computer into the computer platform of an application server. Thus, another type of media that can bear the software elements includes optical, electrical and electromagnetic waves, such as used across physical interfaces between local devices, through wired and optical landline networks and over various air-links. The physical elements that carry such waves, such as wired or wireless links, optical links or the like, also can be considered as media bearing the software. As used herein, unless restricted to non-transitory, tangible “storage” media, terms such as computer or

machine “readable medium” refer to any medium that participates in providing instructions to a processor for execution.

[00121] Hence, a machine readable medium, such as computer-executable code, may take many forms, including but not limited to, a tangible storage medium, a carrier wave medium, or physical transmission medium. Non-volatile storage media include, for example, optical or magnetic disks, such as any of the storage devices in any computer(s) or the like, such as can be used to implement the databases, etc. shown in the drawings. Volatile storage media include dynamic memory, such as main memory of such a computer platform. Tangible transmission media include coaxial cables; copper wire and fiber optics, including the wires that comprise a bus within a computer system.

[00122] Carrier-wave transmission media may take the form of electric or electromagnetic signals, or acoustic or light waves such as those generated during radio frequency (RF) and infrared (IR) data communications. Common forms of computer-readable media therefore include for example: a floppy disk, a flexible disk, hard disk, magnetic tape, any other magnetic medium, a CD-ROM, DVD or DVD-ROM, any other optical medium, punch cards paper tape, any other physical storage medium with patterns of holes, a RAM, a ROM, a PROM and EPROM, a FLASH-EPROM, any other memory chip or cartridge, a carrier wave transporting data or instructions, cables or links transporting such a carrier wave, or any other medium from which a computer may read programming code and/or data. Many of these forms of computer readable media can be involved in carrying one or more sequences of one or more instructions to a processor for execution.

[00123] The computer system can include or be in communication with an electronic display that comprises a user interface (UI) for providing, for example, a current stage of processing or assaying of a sample (e.g., a particular step, such as a lysis step, or sequencing step that is being performed). Inputs are received by the computer system from one or more measurement. Examples of UIs include, without limitation, a graphical user interface (GUI) and web-based user interface. The algorithm can, for example, process and/or assay a sample, perform sequencing analysis, measure sets of values representative of classes of molecules, identify sets of features and feature vectors from assay data, process feature vectors using a machine learning

model to obtain output classifications, and train a machine learning model (e.g., iteratively search for optimal values of parameters of the machine learning model).

[00124] In some embodiments, systems capable of executing one or more algorithms, e.g., laptops, desktops, iPads, mobile devices etc., for determining changes in cfDNA mutation profiles, frequency of mutations and/or fragmentation profiles classifies the subject as a cancer patient based on the cfDNA mutation profiles, frequency of mutations and/or fragmentation for the subject. These systems further execute machine learning algorithms that can be used to generate models such as, for example, high-risk populations and low-risk general populations (a penalized logistic regression with the Mathios *et al.* (Mathios D, Johansen JS, Cristiano S, Medina JE, Phallen J, Larsen KR, *et al.* Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat Commun* 2021;12(1):5060) features as well as coverage from transcription factor binding sites. These models can be trained on the subject cohort with 5-fold cross validation with 10 repeats, and scores for each sample are calculated by the mean across repeats and evaluated using AUC-ROC. For example, the first model used the high-risk non-cancer and HCC patients while the second used the non-cancer individuals without liver pathology. The locked high-risk model trained on the cohort was applied to a second and different cohort to generate cancer predictions on an external validation set. A “class label” can be applied to each sample indicating the classification of the sample for any number of input features. For example, the class labels for the set of cohorts could indicate the identity of cfDNA mutation profiles, frequency of mutations and/or fragmentation profiles based on genomic location etc. The resulting training sets are provided to machine learning unit, such as a neural network or a support vector machine. Using the training set, the machine learning unit may generate a model to classify the sample according to the cfDNA mutation profiles, frequency of mutations and/or fragmentation profile.

[00125] In some embodiments, a method is provided for creating a trained classifier, comprising the steps of: (a) providing a plurality of different classes, wherein each class represents a set of subjects with a shared characteristic (e.g. from one or more cohorts); (b) providing a multi-parametric model representative of the cell-free DNA molecules from each of a plurality of samples belonging to each of the classes, thereby providing a training data set; and (c) training a learning algorithm on the training data set to create one or more trained classifiers, wherein each trained classifier classifies a test sample into one or more of the plurality of classes.

[00126] As an example, a trained classifier may use a learning algorithm selected from the group consisting of: a random forest, a neural network, a support vector machine, and a linear classifier. Each of the plurality of different classes may be selected from the group consisting of healthy, breast cancer, colon cancer, lung cancer, pancreatic cancer, prostate cancer, ovarian cancer, melanoma, and liver cancer.

[00127] A trained classifier may be applied to a method of classifying a sample from a subject. This method of classifying may comprise: (a) providing a multi-parametric model representative of the cell-free DNA molecules from a test sample from the subject; and (b) classifying the test sample using a trained classifier. After the test sample is classified into one or more classes, a therapeutic intervention on the subject can be performed based on the classification of the sample.

[00128] In some embodiments, training sets are provided to a machine learning unit, such as a neural network or a support vector machine. Using the training set, the machine learning unit may generate a model to classify the sample according to a treatment response to one or more therapeutic inventions. This is also referred to as “calling”. The model developed may employ information from any part of a test vector.

[00129] In general, machine learning can be used to reduce a set of data generated from all (primary sample/analytes/test) combinations into an optimal predictive set of features, e.g., which satisfy specified criteria. In various examples statistical learning, and/or regression analysis can be applied. Simple to complex and small to large models making a variety of modeling assumptions can be applied to the data in a cross-validation paradigm. Simple to complex includes considerations of linearity to non-linearity and non-hierarchical to hierarchical representations of the features. Small to large models includes considerations of the size of basis vector space to project the data onto as well as the number of interactions between features that are included in the modelling process.

[00130] Machine learning techniques can be used to assess the commercial testing modalities most optimal for cost/performance/commercial reach as defined in the initial question. A threshold check can be performed: If the method applied to a hold-out dataset that was not used in cross validation surpasses the initialized constraints, then the assay is locked, and production initiated. For example, a threshold for assay performance may include a desired minimum

accuracy, positive predictive value (PPV), negative predictive value (NPV), clinical sensitivity, clinical specificity, area under the curve (AUC), or a combination thereof. For example, a desired minimum accuracy, PPV, NPV, clinical sensitivity, clinical specificity, or combination thereof may be at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 81%, at least about 82%, at least about 83%, at least about 84%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, or at least about 99%. As another example, a desired minimum AUC may be at least about 0.50, at least about 0.55, at least about 0.60, at least about 0.65, at least about 0.70, at least about 0.75, at least about 0.80, at least about 0.81, at least about 0.82, at least about 0.83, at least about 0.84, at least about 0.85, at least about 0.86, at least about 0.87, at least about 0.88, at least about 0.89, at least about 0.90, at least about 0.91, at least about 0.92, at least about 0.93, at least about 0.94, at least about 0.95, at least about 0.96, at least about 0.97, at least about 0.98, or at least about 0.99. A subset of assays may be selected from a set of assays to be performed on a given sample based on the total cost of performing the subset of assays, subject to the threshold for assay performance, such as desired minimum accuracy, positive predictive value (PPV), negative predictive value (NPV), clinical sensitivity, clinical specificity, area under the curve (AUC), and a combination thereof. If the thresholds are not met, then the assay engineering procedure can loop back to either the constraint setting for possible relaxation or to the wet lab to change the parameters in which data was acquired. Given the clinical question, biological constraints, budget, lab machines, etc., can constrain the problem.

[00131] In certain embodiments, the computer processing of a machine learning technique can include method(s) of statistics, mathematics, biology, or any combination thereof. In various examples, any one of the computer processing methods can include a dimension reduction method, logistic regression, dimension reduction, principal component analysis, autoencoders, singular value decomposition, Fourier bases, singular value decomposition, wavelets, discriminant analysis, support vector machine, tree-based methods, random forest, gradient boost tree, logistic regression, matrix factorization, network clustering, statistical testing and neural network.

[00132] In certain embodiments, the computer processing of a machine learning technique can include logistic regression, multiple linear regression (MLR), dimension reduction, partial least squares (PLS) regression, principal component regression, autoencoders, variational autoencoders, singular value decomposition, Fourier bases, wavelets, discriminant analysis, support vector machine, decision tree, classification and regression trees (CART), tree-based methods, random forest, gradient boost tree, logistic regression, matrix factorization, multidimensional scaling (MDS), dimensionality reduction methods, t-distributed stochastic neighbor embedding (t-SNE), multilayer perceptron (MLP), network clustering, neuro-fuzzy, neural networks (shallow and deep), artificial neural networks, Pearson product-moment correlation coefficient, Spearman's rank correlation coefficient, Kendall tau rank correlation coefficient, or any combination thereof. In some examples, the computer processing method is a supervised machine learning method including, for example, a regression, support vector machine, tree-based method, and neural network. In some examples, the computer processing method is an unsupervised machine learning method including, for example, clustering, network, principal component analysis, and matrix factorization.

[00133] For supervised learning, training samples (e.g., in thousands) can include measured data (e.g., of various analytes) and known labels, which may be determined via other time-consuming processes, such as imaging of the subject and analysis by a trained practitioner. Example labels can include classification of a subject, e.g., discrete classification of whether a subject has cancer or not or continuous classifications providing a probability (e.g., a risk or a score) of a discrete value. A learning module can optimize parameters of a model such that a quality metric (e.g., accuracy of prediction to known label) is achieved with one or more specified criteria. Determining a quality metric can be implemented for any arbitrary function including the set of all risk, loss, utility, and decision functions. A gradient can be used in conjunction with a learning step (e.g., a measure of how much the parameters of the model should be updated for a given time step of the optimization process).

[00134] As described above, examples can be used for a variety of purposes. For example, plasma (or other sample) can be collected from subjects symptomatic with a condition (e.g., known to have the condition) and healthy subjects. Genetic data (e.g., cfDNA) can be acquired analyzed to obtain a variety of different features, which can include features based on a genome wide

analysis. These features can form a feature space that is searched, stretched, rotated, translated, and linearly or non-linearly transformed to generate an accurate machine learning model, which can differentiate between healthy subjects and subjects with the condition (e.g., identify a disease or non-disease status of a subject). Output derived from this data and model (which may include probabilities of the condition, stages (levels) of the condition, or other values), can be used to generate another model that can be used to recommend further procedures, e.g., recommend a biopsy or keep monitoring the subject condition.

[00135] In some embodiments, DNA from a population of several individuals can be analyzed by a set of multiplexed arrays. The data for each multiplexed array may be self-normalized using the information contained in that specific array. This normalization algorithm may adjust for nominal intensity variations observed in the two-color channels, background differences between the channels, and possible crosstalk between the dyes. The behavior of each base position may then be modeled using a clustering algorithm that incorporates several biological heuristics on mutation profiles, frequency of mutations and/or fragmentation profiles. In cases where few cfDNA fragments are observed (e.g., due to low minor-allele frequency), locations and shapes of the missing sequences may be estimated using neural networks. Depending on the profiles and percent sequence identity, a statistical score may be devised (a Training score). A score such as GenCall Score is designed to mimic evaluations made by a human expert's visual and cognitive systems. In addition, it has been evolved using the genotyping data from top and bottom strands. This score may be combined with several penalty terms (e.g., low intensity, mismatch between existing and predicted cfDNA fragments) in order to make up the Training score. The Training score is saved for use by the calling algorithm.

[00136] To call a therapeutic response, a calling algorithm may take the genetic information and treatment responses of a plurality of individuals having a disease or condition. The data may first be normalized (using the same procedure as for the clustering algorithm). The calling operation (classification) may be performed using, for example, a Bayesian model. The score for each call's Call Score can be the product of a Training Score and a data-to-model fit score. After scoring all the treatment responses, the application may compute a composite score.

[00137] In some embodiments, a training dataset comprises clinical data selected from the group consisting of cancer stage, type of surgical procedure, age, tumor grading, depth of tumor

infiltration, occurrence of post-operative complications, and the presence of venous invasion. In some embodiments, the training dataset is pre-processed, comprising transforming the provided data into class-conditional probabilities.

[00138] Another embodiment uses machine learning techniques to train a statistical classifier, specifically a support vector machine, for each cancer stage category based on word occurrences in a corpus of histology reports for each patient. New reports can then be classified according to the most likely stage, facilitating the collection and analysis of population staging data.

[00139] In some embodiments, a machine learning algorithm is selected from the group consisting of: a supervised or unsupervised learning algorithm selected from support vector machine, random forest, nearest neighbor analysis, linear regression, binary decision tree, discriminant analyses, logistic classifier, and cluster analysis.

[00140] In general, a system can comprise a report generator for reporting on cancer test results and treatment options. The report generator system can be a central data processing system configured to establish communications directly with: a remote data site or laboratory, a medical practice/healthcare provider (treating professional) and/or a patient/subject through communication links. The laboratory can be medical laboratory, diagnostic laboratory, medical facility, medical practice, point-of-care testing device, or any other remote data site capable of generating subject clinical information. Subject clinical information includes but it is not limited to laboratory test data, X-ray data, examination and diagnosis. The healthcare provider or practice 26 includes medical services providers, such as doctors, nurses, home health aides, technicians and physician's assistants, and the practice is any medical care facility staffed with healthcare providers. In certain instances, the healthcare provider/practice is also a remote data site. In a cancer treatment embodiment, the subject may be afflicted with cancer, among others.

[00141] Other clinical information for a cancer subject includes the results of laboratory tests, imaging or medical procedure directed towards the specific cancer that one of ordinary skill in the art can readily identify. The list of appropriate sources of clinical information for cancer includes but it is not limited to: CT scan, MRI scan, ultrasound scan, bone scan, PET Scan, bone marrow test, barium X-ray, endoscopy, lymphangiogram, IVU (Intravenous urogram) or IVP (IV

pyelogram), lumbar puncture, cystoscopy, immunological tests (anti-malignin antibody screen), and cancer marker tests.

[00142] The subject clinical information may be obtained from the laboratory manually or automatically. For simplicity of the system the information is obtained automatically at predetermined or regular time intervals. A regular time interval refers to a time interval at which the collection of the laboratory data is carried out automatically by the methods and systems described herein based on a measurement of time such as hours, days, weeks, months, years etc. In one embodiment of the invention, the collection of data and processing is carried out at least once a day. In one embodiment, the transfer and collection of data is carried out once every month, biweekly, or once a week, or once every couple of days. Alternatively, the retrieval of information may be carried out at predetermined but not regular time intervals. For instance, a first retrieval step may occur after one week and a second retrieval step may occur after one month. The transfer and collection of data can be customized according to the nature of the disorder that is being managed and the frequency of required testing and medical examinations of the subjects.

[00143] In certain embodiments, a genetic report is generated from a subject's sample, e.g. cfDNA. The polynucleotides in a sample can be sequenced, e.g., whole genome sequencing, NGS sequencing, producing a plurality of sequence reads. In some embodiments, genetic information comprises variables defining the genomic organization of cancer cells or the genomic organization of single disseminated cancer cells. In some embodiments, the genetic information comprises sequence or abundance data from one or more genetic loci in cell-free DNA from the individuals.

[00144] cfDNA genetic information is processed (72). Genetic variants can also be identified. Genetic variants include sequence variants, copy number variants and nucleotide modification variants. A sequence variant is a variation in a genetic nucleotide sequence. A copy number variant is a deviation from wild type in the number of copies of a portion of a genome. Genetic variants include, for example, single nucleotide variations (SNPs), insertions, deletions, inversions, transversions, translocations, gene fusions, chromosome fusions, gene truncations, copy number variations (e.g., aneuploidy, partial aneuploidy, polyploidy, gene amplification), abnormal changes in nucleic acid chemical modifications, abnormal changes in epigenetic patterns and abnormal changes in nucleic acid methylation. The process then determines the frequency of genetic variants in the sample containing the genetic material. Since this process is noisy, the

process separates information from noise (73). The sensitivity of detecting genetic variants can be increased by increasing read depth of polynucleotides (e.g., by sequencing to a greater read depth at in a sample from a subject at two or more time points).

[00145] To increase the diagnosis confidence, a plurality of measurements can be taken. Or alternatively using measurements at a plurality of time points (e.g., 2, 3, 4, 5, 6, 7, 8, 9, 10, or more time points) to determine whether cancer is advancing, in remission or stabilized. The diagnostic confidence can be used to identify disease states. For example, cell free polynucleotides taken from a subject can include polynucleotides derived from normal cells, as well as polynucleotides derived from diseased cells, such as cancer cells. Polynucleotides from cancer cells may bear genetic variants, such as somatic cell mutations and copy number variants. When cell free polynucleotides from a sample from a subject are sequenced, and cfDNA mutation profiles, frequency of mutations and/or fragmentation profiles can be produced as described in the examples section which follows.

[00146] Numerous cancers may be detected using the methods and systems described herein. Cancers cells, as most cells, can be characterized by a rate of turnover, in which old cells die and replaced by newer cells. Generally dead cells, in contact with vasculature in a given subject, may release DNA or fragments of DNA into the blood stream. This is also true of cancer cells during various stages of the disease. Cancer cells may also be characterized, dependent on the stage of the disease, by various genetic aberrations such as copy number variation as well as mutations. This phenomenon may be used to detect the presence or absence of cancers individuals using the methods and systems described herein.

[00147] In the early detection of cancers, any of the systems or methods herein described, including mutation detection or copy number variation detection may be utilized to detect cancers. These system and methods may be used to detect any number of genetic aberrations that may cause or result from cancers. These may include but are not limited to cfDNA mutation profiles, frequency of mutations, cfDNA fragmentation profiles, mutations, mutations, indels, copy number variations, transversions, translocations, inversion, deletions, aneuploidy, partial aneuploidy, polyploidy, chromosomal instability, chromosomal structure alterations, gene fusions, chromosome fusions, gene truncations, gene amplification, gene duplications, chromosomal lesions, DNA lesions, abnormal changes in nucleic acid chemical modifications, abnormal changes in epigenetic patterns, abnormal changes in nucleic acid methylation infection and cancer.

[00148] Additionally, the systems and methods described herein may also be used to help characterize certain cancers. Genetic data produced from the system and methods of this disclosure may allow practitioners to help better characterize a specific form of cancer. Often times, cancers are heterogeneous in both composition and staging. Genetic profile data may allow characterization of specific sub-types of cancer that may be important in the diagnosis or treatment of that specific sub-type. This information may also provide a subject or practitioner clues regarding the prognosis of a specific type of cancer.

[00149] The systems and methods provided herein may be used to monitor already known cancers, or other diseases in a particular subject. This may allow either a subject or practitioner to adapt treatment options in accord with the progress of the disease. In this example, the systems and methods described herein may be used to construct genetic cfDNA mutation profiles, frequency of mutations and/or fragmentation profiles of a particular subject of the course of the disease. In some instances, cancers can progress, becoming more aggressive and genetically unstable. In other examples, cancers may remain benign, inactive or dormant. The system and methods of this disclosure may be useful in determining disease progression.

[00150] Further, the systems and methods described herein may be useful in determining the efficacy of a particular treatment option. In one example, certain treatment options may be correlated with genetic cfDNA mutation profiles, frequency of mutations and/or fragmentation profiles of cancers over time. This correlation may be useful in selecting a therapy. Additionally, if a cancer is observed to be in remission after treatment, the systems and methods described herein may be useful in monitoring residual disease or recurrence of disease.

[00151] Further, the methods of the disclosure may be used to characterize the heterogeneity of an abnormal condition in a subject, the method comprising generating a cfDNA mutation profile, frequency of mutations and/or fragmentation profile of extracellular polynucleotides in the subject, wherein the cfDNA mutation profile comprises a plurality of data resulting from profile variation and mutation analyses. In some cases, including but not limited to cancer, a disease may be heterogeneous. Disease cells may not be identical. In the example of cancer, some tumors are known to comprise different types of tumor cells, some cells in different stages of the cancer. In other examples, heterogeneity may comprise multiple foci of disease. Again, in the example of

cancer, there may be multiple tumor foci, perhaps where one or more foci are the result of metastases that have spread from a primary site (also known as distant metastases).

[00152] The methods of this disclosure may be used to generate a profile, fingerprint, or set of data that is a summation of genetic information derived from different cells in a heterogeneous disease. This set of data may comprise copy number variation and mutation analyses alone or in combination.

[00153] Further, these reports are submitted and accessed electronically via the internet. Analysis of data occurs at a site other than the location of the subject. The report is generated and transmitted to the subject's location. Via an internet enabled computer, the subject accesses the reports reflecting his tumor burden.

[00154] The annotated information can be used by a health care provider to select other drug treatment options and/or provide information about drug treatment options to an insurance company. The method can include annotating the drug treatment options for a condition in, for example, the NCCN Clinical Practice Guidelines in Oncology™ or the American Society of Clinical Oncology (ASCO) clinical practice guidelines.

[00155] Reports are generated, mapping genome positions and cfDNA mutation profile variation for the subject with cancer. These reports, in comparison to other profiles of subjects with known outcomes, can indicate that a particular cancer is aggressive and resistant to treatment. The subject is monitored for a period and retested. If at the end of the period, the cfDNA mutation profiles, frequency of mutations and/or fragmentation variation profile does not vary, this may indicate that the current treatment is not working. A comparison is done with cfDNA mutation profiles of other subjects. For example, if it is determined that a change in cfDNA mutation variation indicates that the cancer is advancing, then the original treatment regimen as prescribed is no longer treating the cancer and a new treatment is prescribed.

[00156] In certain embodiments, the system receives genetic information from a DNA sequencer. The process then determines specific cfDNA alterations and frequencies thereof. These reports are submitted and accessed electronically via the internet. Analysis of data occurs at a site other than the location of the subject. The report is generated and transmitted to the subject's location. Via an internet enabled computer, the subject accesses the reports reflecting his tumor burden.

[00157] While temporal information can be used to enhance the information for cfDNA mutation profiles and frequency of mutations, other consensus methods can be applied. In other embodiments, the historical comparison can be used in conjunction with other consensus cfDNA mutation profiles, frequency of mutations and/or fragmentation profiles. Consensus cfDNA mutation profiles and frequency of mutations can be normalized against control samples. Measures of molecules mapping to reference sequences can also be compared across a genome to identify areas in the genome in which cfDNA mutation profiles and frequency of mutations varies, or remains the same. Consensus methods include, for example, linear or non-linear methods of building consensus cfDNA mutation profiles and frequency of mutations (such as voting, averaging, statistical, maximum a posteriori or maximum likelihood detection, dynamic programming, Bayesian, hidden Markov or support vector machine methods, etc.) derived from digital communication theory, information theory, or bioinformatics. After the sequence read coverage has been determined, a stochastic modeling algorithm is applied to convert the normalized nucleic acid sequence read coverage for each window region to the discrete copy number states. In some cases, this algorithm may comprise one or more of the following: Hidden Markov Model, dynamic programming, support vector machine, Bayesian network, trellis decoding, Viterbi decoding, expectation maximization, Kalman filtering methodologies and neural networks.

[00158] Artificial neural networks (NNets) mimic networks of “neurons” based on the neural structure of the brain. They process records one at a time, or in a batch mode, and “learn” by comparing their classification of the record (which, at the outset, is largely arbitrary) with the known actual classification of the record. In MLP-NNets, the errors from the initial classification of the first record is fed back into the network, and are used to modify the network's algorithm the second time around, and so on for many iterations. The neural networks use an iterative learning process in which data cases (rows) are presented to the network one at a time, and the weights associated with the input values are adjusted each time.

[00159] After all cases are presented, the process often starts over again. During this learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of input samples. Neural network learning is also referred to as “connectionist learning,” due to connections between the units. Advantages of neural networks include their high tolerance

to noisy data, as well as their ability to classify patterns on which they have not been trained. One neural network algorithm is back-propagation algorithm, such as Levenberg-Marquadt. Once a network has been structured for a particular application, that network is ready to be trained. To start this process, the initial weights are chosen randomly. Then the training, or learning, begins.

[00160] The network processes the records in the training data one at a time, using the weights and functions in the hidden layers, then compares the resulting outputs against the desired outputs. Errors are then propagated back through the system, causing the system to adjust the weights for application to the next record to be processed. This process occurs over and over as the weights are continually tweaked. During the training of a network the same set of data is processed many times as the connection weights are continually refined.

[00161] In an embodiment, the training step of the machine learning unit on the training data set may generate one or more classification models for applying to a test sample. These classification models may be applied to a test sample to predict the response of a subject to a therapeutic intervention.

[00162] Comparison of sequence coverage to a control sample or reference sequence may aid in normalization across windows. In this embodiment, cell free DNAs are extracted and isolated from a readily accessible bodily fluid such as blood. For example, cell free DNAs can be extracted using a variety of methods known in the art, including but not limited to isopropanol precipitation and/or silica based purification. Cell free DNAs may be extracted from any number of subjects, such as subjects without cancer, subjects at risk for cancer, or subjects known to have cancer (e.g. through other means).

[00163] Following the isolation/extraction step, any of a number of different sequencing operations may be performed on the cell free polynucleotide sample. Samples may be processed before sequencing with one or more reagents (e.g., enzymes, unique identifiers (e.g., barcodes), probes, etc.). In some cases, if the sample is processed with a unique identifier such as a barcode, the samples or fragments of samples may be tagged individually or in subgroups with the unique identifier. The tagged sample may then be used in a downstream application such as a sequencing reaction by which individual molecules may be tracked to parent molecules.

[00164] The cell free polynucleotides can be tagged or tracked in order to permit subsequent identification and origin of the particular polynucleotide. The assignment of an identifier (e.g., a

barcode) to individual or subgroups of polynucleotides may allow for a unique identity to be assigned to individual sequences or fragments of sequences. This may allow acquisition of data from individual samples and is not limited to averages of samples. In some examples, nucleic acids or other molecules derived from a single strand may share a common tag or identifier and therefore may be later identified as being derived from that strand. Similarly, all of the fragments from a single strand of nucleic acid may be tagged with the same identifier or tag, thereby permitting subsequent identification of fragments from the parent strand. In other cases, gene expression products (e.g., mRNA) may be tagged in order to quantify expression, by which the barcode, or the barcode in combination with sequence to which it is attached can be counted. In still other cases, the systems and methods can be used as a PCR amplification control. In such cases, multiple amplification products from a PCR reaction can be tagged with the same tag or identifier. If the products are later sequenced and demonstrate sequence differences, differences among products with the same identifier can then be attributed to PCR error. Additionally, individual sequences may be identified based upon characteristics of sequence data for the read themselves. For example, the detection of unique sequence data at the beginning (start) and end (stop) portions of individual sequencing reads may be used, alone or in combination, with the length, or number of base pairs of each sequence read unique sequence to assign unique identities to individual molecules. Fragments from a single strand of nucleic acid, having been assigned a unique identity, may thereby permit subsequent identification of fragments from the parent strand. This can be used in conjunction with bottlenecking the initial starting genetic material to limit diversity.

[00165] Generally, the methods and systems provided herein are useful for preparation of cell free polynucleotide sequences to a down-stream application sequencing reaction. Often, a sequencing method is next generation sequencing (NGS), classic Sanger sequencing, whole-genome bisulfite sequencing (WGSB), small-RNA sequencing, low-coverage Whole-Genome Sequencing (lcWGS), etc.

[00166] As used herein, the term “sequencing” refers to any of a number of technologies used to determine the sequence of a biomolecule, e.g., a nucleic acid such as DNA or RNA. Exemplary sequencing methods include, but are not limited to, targeted sequencing, single molecule real-time sequencing, exon sequencing, electron microscopy-based sequencing, panel sequencing, transistor-mediated sequencing, direct sequencing, random shotgun sequencing,

Sanger dideoxy termination sequencing, whole-genome sequencing, sequencing by hybridization, pyrosequencing, capillary electrophoresis, gel electrophoresis, duplex sequencing, cycle sequencing, single-base extension sequencing, solid-phase sequencing, high-throughput sequencing, massively parallel signature sequencing, emulsion PCR, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, sequencing by reversible dye terminator, paired-end sequencing, near-term sequencing, exonuclease sequencing, sequencing by ligation, short-read sequencing, single-molecule sequencing, sequencing-by-synthesis, real-time sequencing, reverse-terminator sequencing, nanopore sequencing, 454 sequencing, Solexa Genome Analyzer sequencing, SOLiD™ sequencing, MS-PET sequencing, and a combination thereof. In some embodiments, sequencing can be performed by a gene analyzer such as, for example, gene analyzers commercially available from Illumina or Applied Biosystems. In some embodiments, the sequencing method can be massively parallel sequencing, that is, simultaneously (or in rapid succession) sequencing any of at least 100, 1000, 10,000, 100,000, 1 million, 10 million, 100 million, or 1 billion polynucleotide molecules.

[00167] After sequencing, reads are assigned a quality score. A quality score may be a representation of reads that indicates whether those reads may be useful in subsequent analysis based on a threshold. In some cases, some reads are not of sufficient quality or length to perform the subsequent mapping step. Sequencing reads with a quality score at least 90%, 95%, 99%, 99.9%, 99.99% or 99.999% may be filtered out of the data set. In other cases, sequencing reads assigned a quality score at least 90%, 95%, 99%, 99.9%, 99.99% or 99.999% may be filtered out of the data set. The genomic fragment reads that meet a specified quality score threshold are mapped to a reference genome, or a reference sequence that is known not to contain mutations. After mapping alignment, sequence reads are assigned a mapping score. A mapping score may be a representation of reads mapped back to the reference sequence indicating whether each position is or is not uniquely mappable. In instances, reads may be sequences unrelated to mutation analysis. For example, some sequence reads may originate from contaminant polynucleotides. Sequencing reads with a mapping score at least 90%, 95%, 99%, 99.9%, 99.99% or 99.999% may be filtered out of the data set. In other cases, sequencing reads assigned a mapping score less than 90%, 95%, 99%, 99.9%, 99.99% or 99.999% may be filtered out of the data set. For each mappable base, bases that do not meet the minimum threshold for mappability, or low quality bases, may be replaced by the corresponding bases as found in the reference sequence.

[00168] Numerous cancers may be detected using the methods and systems described herein. Cancers cells, as most cells, can be characterized by a rate of turnover, in which old cells die and replaced by newer cells. Generally dead cells, in contact with vasculature in a given subject, may release DNA or fragments of DNA into the blood stream. This is also true of cancer cells during various stages of the disease. Cancer cells may also be characterized, dependent on the stage of the disease, by various genetic aberrations such as copy number variation as well as mutations. This phenomenon may be used to detect the presence or absence of cancers individuals using the methods and systems described herein.

[00169] The types and number of cancers that may be detected may include but are not limited to blood cancers, brain cancers, lung cancers, skin cancers, nose cancers, throat cancers, liver cancers, bone cancers, lymphomas, pancreatic cancers, skin cancers, bowel cancers, rectal cancers, thyroid cancers, bladder cancers, kidney cancers, mouth cancers, stomach cancers, solid state tumors, heterogeneous tumors, homogenous tumors and the like.

[00170] Additionally, the systems and methods described herein may also be used to help characterize certain cancers. Genetic data produced from the system and methods of this disclosure may allow practitioners to help better characterize a specific form of cancer. Often times, cancers are heterogeneous in both composition and staging. Genetic profile data may allow characterization of specific sub-types of cancer that may be important in the diagnosis or treatment of that specific sub-type. This information may also provide a subject or practitioner clues regarding the prognosis of a specific type of cancer.

[00171] The systems and methods provided herein may be used to monitor already known cancers, or other diseases in a particular subject. This may allow either a subject or practitioner to adapt treatment options in accord with the progress of the disease. In this example, the systems and methods described herein may be used to construct genetic profiles of a particular subject of the course of the disease. In some instances, cancers can progress, becoming more aggressive and genetically unstable. In other examples, cancers may remain benign, inactive or dormant. The system and methods of this disclosure may be useful in determining disease progression.

[00172] Further, the systems and methods described herein may be useful in determining the efficacy of a particular treatment option. In one example, successful treatment options may actually increase the amount of copy number variation or mutations detected in subject's blood if

the treatment is successful as more cancers may die and shed DNA. In other examples, this may not occur. In another example, perhaps certain treatment options may be correlated with genetic profiles of cancers over time. This correlation may be useful in selecting a therapy. Additionally, if a cancer is observed to be in remission after treatment, the systems and methods described herein may be useful in monitoring residual disease or recurrence of disease.

[00173] The data is sent over a direct connection or over the internet to a computer for processing. The data processing aspects of the system can be implemented in digital electronic circuitry, or in computer hardware, firmware, software, or in combinations of them. Data processing apparatus of the invention can be implemented in a computer program product tangibly embodied in a machine-readable storage device for execution by a programmable processor; and data processing method steps of the invention can be performed by a programmable processor executing a program of instructions to perform functions of the invention by operating on input data and generating output. The data processing aspects of the invention can be implemented advantageously in one or more computer programs that are executable on a programmable system including at least one programmable processor coupled to receive data and instructions from and to transmit data and instructions to a data storage system, at least one input device, and at least one output device. Each computer program can be implemented in a high-level procedural or object-oriented programming language, or in assembly or machine language, if desired; and, in any case, the language can be a compiled or interpreted language. Suitable processors include, by way of example, both general and special purpose microprocessors. Generally, a processor will receive instructions and data from a read-only memory and/or a random access memory. Storage devices suitable for tangibly embodying computer program instructions and data include all forms of nonvolatile memory, including by way of example semiconductor memory devices, such as EPROM, EEPROM, and flash memory devices; magnetic disks such as internal hard disks and removable disks; magneto-optical disks; and CD-ROM disks. Any of the foregoing can be supplemented by, or incorporated in, ASICs (application-specific integrated circuits).

[00174] To provide for interaction with a user, the methods can be implemented using a computer system having a display device such as a monitor or LCD (liquid crystal display) screen for displaying information to the user and input devices by which the user can provide input to the computer system such as a keyboard, a two-dimensional pointing device such as a mouse or a

trackball, or a three-dimensional pointing device such as a data glove or a gyroscopic mouse. The computer system can be programmed to provide a graphical user interface through which computer programs interact with users. The computer system can be programmed to provide a virtual reality, three-dimensional display interface.

EXAMPLES

[00175] *Example 1: Single Molecule Genome-Wide Mutation And Fragmentation Profiles Of Cell-Free DNA For Noninvasive Detection Of Lung Cancer*

[00176] It was considered whether identifying somatic sequence changes across the genome could allow for detection of an increased number of tumor derived cfDNA changes and increase the ability to detect early stage disease. A tumor genome contains thousands of somatic changes^{19,20} and knowledge of such alterations from tumor tissue have been used to guide targeted analyses in the circulation during therapy^{21,22}. In principle, if such genome-wide changes could be identified directly in cfDNA without knowledge of the alterations in the tumor, they could be useful for early cancer detection. However, such an approach would require the ability to efficiently detect somatic changes and to distinguish these from the multitude of other non-tumor derived alterations.

[00177] To address these challenges, an approach was developed herein, called GENome-wide Mutational Incidence for NonInvasive detection of cancer (GEMINI), that could identify a much larger number of tumor-derived alterations in cfDNA for cancer detection (FIG. 1). This method was applied to analyze tissue and cfDNA samples from multiple patient cohorts (FIG. 6). The method is based on sequencing individual cfDNA molecules to estimate the mutation frequency and type of alteration across the genome using non-overlapping bins ranging in size from thousands to millions of bases. For each individual, the mutation type and frequency in genomic regions more commonly altered in cancer is compared to the profile from regions more frequently mutated in normal cfDNA to determine multiregional differences in mutation profiles. In this way, the GEMINI approach enriches for likely somatic mutations while taking into account individual variability in overall background changes.

[0178] Results

[0179] To develop this method, whole genome sequences of cancers were examined from 2,511 individuals across 25 different cancers from the Pan-Cancer Analysis of Whole Genomes

(PCAWG) study^{25,26}, identifying distinct frequencies of somatic mutations across the genome in different tumor types (FIG. 7; Supplementary Table1). As an example, analysis of lung tumor and matched normal tissue genomes from 65 individuals with smoking exposure revealed that the cancers had an average of 52,209 (range 6,031 to 193,539) *bona fide* somatic mutations per genome. To quantify how many tumor-specific changes would be expected to see in the plasma in these individuals, *in silico* dilution and downsampling experiments were performed (FIG. 8). In these simulations, it was found that all patients would theoretically have a subset of detectable somatic mutations even at a tumor fraction as low as 1:10,000 when analyzed using whole genome sequencing at a coverage of 1x (FIG. 2A).

[0180] Because the majority of mutations detected at low whole genome coverage would be expected to be observed in single DNA molecules (FIG. 2B), rigorous methods were developed to examine the frequency of single molecule somatic mutations in a mixture of germline variants, WBC alterations, and experimental and sequencing artifacts (all considered background changes). Each sequenced molecule was scanned for single nucleotide changes, and after removing common germline variants and unevaluable regions, computed the frequency of putative mutations in high quality reads, defined as the number of variants per million evaluated positions across all the DNA molecules sequenced (Methods). As specific transversions likely related to accumulation of 7,8-dihydro-8-oxoguanine (8-oxo-dG)²⁷ were more highly represented among PCAWG single molecule alterations than other changes and higher than one would expect from analyses of similar transversions at sites of known polymorphisms, these changes were filtered from further consideration when they occurred in certain read combinations (FIGS. 9A, 9B, Methods). These changes were examined in the set of PCAWG lung tumors samples with matched normal blood cells (n=31), as blood cells represent the largest source of cfDNA in non-cancer individuals²⁸. The analyses were focused on the remaining C:G>A:T mutations (hereafter C>A) given their high abundance in tumors of current and former smokers²⁹. As expected, given the high and variable frequency of overall background changes, it was found that C>A mutation frequencies were similar in the tumor compared to normal samples (FIG. 2C), and were only slightly higher even after the filtering steps above and removal of germline variants, where only a small fraction of the tumor alterations was somatic in origin (average=7.5%, range 0.8%-22%) (FIGS. 2D, 10A, 11).

[0181] The high amount and variability of total background changes were investigated among samples, and it was found that these were largely related to sequencing lane- and run-specific artifacts (FIG. 12). It was hypothesized that controlling for overall background rates in a sample-specific manner could improve detection of tumor-derived changes. Previous analyses have shown that mutation rates differ across individual cancer genomes, with regions associated with euchromatin, including expressed genes and early-replicating regions, having a lower mutation rate compared to heterochromatin regions representing unexpressed genes and late-replicating regions^{30,31}. To examine the variation in mutation frequency across the genome, the 31 PCAWG paired samples were analyzed by binning the sequence data containing 3,076,901 mutations into 1,144 non-overlapping 2.5 megabase (Mb) bins and found regions throughout the genome with increased mutation frequencies shared by many tumors (FIGS. 13, 14).

[0182] To evaluate the GEMINI approach for detection of tumor-derived DNA, we identified genomic regions with the highest C>A changes in a training set of cancers and controls and computed the average C>A difference at these regions for patients not represented in the training set (FIG. 15, Methods). Regions enriched for C>A changes were identified in the 31 PCAWG cancers but not normal samples (FIG. 2E) and it was found that background changes were highly correlated in cancer and control regions for each patient sample (Pearson correlation coefficient=0.99, $p<0.0001$) (FIG. 2F), suggesting that subtraction of alteration frequencies between cancer and control regions within a given patient sample would be useful for removing background mutations. In contrast, subtraction of specific mutations observed in the matched normal sample observed in the single-molecule sequencing data was not effective in removing background changes (FIG. 16) because such alterations typically occurred *de novo* and were seen once (FIG. 2G). After background subtraction, the remaining regional mutation frequencies were appreciably higher in tumors compared to normal samples (median of 13.4 compared to 1.3, respectively, $p<0.0001$, Wilcoxon rank sum test), with a high fraction of changes resulting from somatic mutations (average = 80%, range = 31%-100%) (FIG. 2H), and were highly correlated to the frequency of high-confidence somatic C>A changes reported in these samples by the PCAWG consortium (Pearson correlation coefficient = 0.96, $p<0.0001$) (FIG. 2I). The GEMINI approach using C>A regional frequencies was able to distinguish PCAWG cancer from non-cancer samples with high accuracy (AUC = 0.91, 95% CI=0.84-0.99) compared to mutation frequencies alone (AUC=0.64, 95% CI=0.50-0.79) using low coverage whole genome

sequencing (FIGS. 2J, 10A, 10B). The overall approach for filtering background changes using variant quality filters, germline variant removal, and subtraction of regional mutation frequencies from single molecule sequencing resulted in a 1,903-fold enrichment in somatic mutations in these samples.

[0183] To determine if the GEMINI method could be used to noninvasively detect alterations in cfDNA from patients with cancer, the ability of the approach to detect sequence alterations was evaluated in individuals from a prospective lung cancer diagnostic cohort (LUCAS)¹⁸. Low coverage plasma whole genome sequence data (~2x coverage) from the 365 individuals examined in this trial was analyzed, the majority of whom were at high-risk for lung cancer (age 50-80 and smoking history ≥ 20 pack years; Supplementary Table 2) and where blood samples were collected prior to clinical diagnosis. Given the short length of cfDNA fragments¹³, in addition to the filtering steps above and removal of 8-oxo-dG-related changes, the additional filter of restricting analyses to regions with identical sequence calls in overlapping reads in the paired-end library was employed. Requiring a Phred quality score ≥ 30 in both reads would theoretically reduce the error rate of a mutation due to sequencing errors and benefits from a higher degree of overlap for shorter tumor derived cfDNA sequences³², thereby potentially enriching detection of circulating tumor DNA (ctDNA) alterations. Supplementary Table 5 is a summary of cfDNA samples and the genomic analyses of lung cancer patients undergoing targeted therapy.

[0184] Comparison of single molecule sequencing of cfDNA from a subset of the LUCAS cohort and PCAWG tumor tissue samples revealed that the genome regions with increased frequency of cancer-type specific mutations were largely similar between tumor tissue and cfDNA of patients with lung cancer, as well as between tissue and cfDNA of patients with melanoma, and B-cell non-Hodgkin lymphoma (BNHL) (Pearson correlation >0.80 , $p < 0.001$ in all cases) and were located in areas of the genome associated with tissue-specific late replication timing (FIGS. 3A, 3B). Different mutation types among the tumors analyzed contributed to the increased mutation frequencies, including C>A changes in lung cancer, C>T in melanoma, and T>G in lymphoma. It was also found that tumor- and mutation type-specific regional mutation frequencies were related to gene expression³⁰, genome compartmentalization as measured by eigenvector analyses of methylation³³, as well histone 3 lysine 9 tri-methylation (H3K9me3), a

known mark of heterochromatin³⁴, and were consistent between tumor and cfDNA analyses (Pearson correlation >0.80 , $p<0.001$ in all cases) (FIGS. 17A-17C). Individuals without cancer or mutation types or regions not enriched in cancer did not have or were weakly correlated to these characteristics (FIGS. 3B, 17A-17C). Overall, these results suggest that mutation rate variability across the genome in cfDNA is related to chromatin organization and can be leveraged by the GEMINI approach to detect tumor-derived sequence changes in the circulation.

[0185] Using the GEMINI approach, cross-validated regional differences were identified in single molecule mutation frequencies for individuals in the LUCAS cohort. Similar to analyses in PCAWG lung cancers, regional C>A mutation frequencies were preferentially altered in individuals with lung cancer compared to non-cancer individuals ($p<0.0001$, Wilcoxon rank sum test) (FIG. 18) and in contrast to overall C>A frequencies were stable across sequencing lanes (FIG. 19). The regions identified were largely consistent across cross-validation folds and comprised high quality sequences with similar evaluable bases, copy number levels, and mappability but were located at genomic positions associated with different mutational frequencies between individuals with and without lung cancer reflecting the epigenomic characteristics described above (FIGS. 20A-20K, Supplementary Table 6). The identified regional differences in C>A mutation frequencies to CC>AA doublet mutations were further compared because these doublet mutations are enriched in lung cancers of smoking individuals²⁶ and have a very low likelihood of occurring by chance given the requirement of two identical changes occurring in adjacent positions (FIGS. 21A-21F). It was found that the frequency of high quality CC>AA changes was highly correlated with the regional difference in single molecule C>A frequency in both tissue (Spearman's $\rho = 0.62$, $p=0.0002$) and cfDNA samples (Spearman's $\rho = 0.65$, $p<0.0001$) (FIGS. 21E, 21F). These data strongly support the notion that GEMINI mutational frequencies reflect tumor-derived sequence changes in the circulation.

[0186] The regional differences were calibrated in single molecule C>A frequencies to GEMINI scores, reflecting an individual's probability of cancer (Methods). To evaluate whether clinical characteristics may affect genome-wide cfDNA mutation profiles, it was investigated whether non-malignant nodules, sex, age, or the presence of chronic obstructive pulmonary disease (COPD) or autoimmune diseases were associated with GEMINI scores. No difference in the GEMINI scores was observed when comparing non-cancer individuals with and without

benign lesions (median GEMINI score 0.30 vs 0.33, $p=0.94$, Wilcoxon rank sum test) (FIG. 4A), or between men and women ($p=0.14$) (FIGS. 22A-22F). A correlation of the GEMINI score with age ($\rho=-0.15$, $p=0.053$) was not observed, nor was it observed with the levels of the inflammatory markers CRP ($\rho=0.01$, $p=0.89$) or IL-6 ($\rho=-0.07$, $p=0.40$) (FIGS. 22A-22F). Similarly, changes were not observed in the GEMINI score between individuals with or without COPD ($p=0.73$) or autoimmune disease ($p=0.31$) (FIGS. 22A-22F). Together, these analyses show that single molecule mutation frequencies in cfDNA are not considerably affected by demographic characteristics or the presence of acute or chronic inflammatory conditions.

[0187] Next, the relationship between GEMINI scores and cancer stage and histology was evaluated. While the GEMINI score for non-cancer individuals was low (median scores of 0.30 and 0.33 for those with or without benign lesions, respectively), patients with cancer had significantly higher median scores across stages (stage I=0.74, stage II=0.67, stage III=0.76, and stage IV=0.74) ($p<0.001$ for stages I, II, III, or IV, Wilcoxon rank sum test) (Fig. 4a) and histologic subtypes (adenocarcinoma = 0.71, squamous cell carcinoma = 0.72, small cell lung cancer (SCLC) = 0.98) ($p<0.0001$ for all subtypes, Wilcoxon rank sum test) (FIG. 4A). As expected, GEMINI scores were generally related to ctDNA levels, increasing with the tumor fraction estimated by ichorCNA³⁵ ($p<0.0001$, Wilcoxon rank sum test) (FIG. 18A). Higher GEMINI scores in SCLC patients likely reflect the known higher ctDNA fractions in this tumor type³⁶ and patients with lower GEMINI scores (e.g. below 0.5) were more likely to have a NSCLC histology, Supplementary Table7, ($p=0.03$, Fisher's exact test) and lower ichorCNA tumor fraction estimates ($p=0.003$, Wilcoxon rank sum test). A receiver operating characteristic (ROC) curve representing the sensitivity and specificity of the GEMINI approach to identify cancer patients revealed an overall area under the curve (AUC) of 0.85 (95% CI = 0.79-0.91) (FIG. 4D), with high levels of detection across stages and subtypes (FIGS. 4F, 24A, 24B).

[0188] The fixed GEMINI model was used to evaluate samples from seven patients who did not have cancer at the time of blood collection but were subsequently diagnosed with lung cancer. These individuals had median GEMINI scores of 0.78, significantly higher than those of non-cancer individuals ($p = 0.0005$, Wilcoxon rank sum test) (FIG. 4B). Six of seven of these individuals had a score above the threshold at an 80% specificity, with the time to lung cancer diagnosis ranging from 231 to 1868 days, providing evidence that abnormalities in cfDNA

mutational profiles would be useful for cancer detection years earlier than standard diagnoses. Of these patients, 5 were diagnosed with NSCLC (two patients had stage I disease, one patient had stage III disease, and stage information was not available for the other two patients), one patient was diagnosed with SCLC (unknown stage), and the other patient for which we do not have stage or histology information died within a few months of their diagnosis. The patient who was not detected by GEMINI had the longest time from blood draw to diagnosis (1,954 days). Interestingly, at the time of the initial blood draw, cancer was not suspected for four of these patients based on CT imaging and no biopsy was performed. For the remaining three patients, there was suspicion of cancer based on CT imaging and the patients underwent biopsy, however their pathology report indicated a benign lung nodule, highlighting limitations of current diagnostic approaches.

[0189] It was next examined whether GEMINI mutational profiles could be combined with genome-wide fragmentation features used by the DELFI approach as it was hypothesized that these methods measured complementary cfDNA characteristics and could be used to improve the ability to detect individuals with early stage lung cancer. The GEMINI and DELFI scores were integrated into a combined score to evaluate the predictive accuracy relative to these features used in isolation (Methods). While GEMINI and DELFI scores were positively correlated (Spearman's $\rho = 0.50$, $p < 0.0001$), several samples missed by either approach in isolation were detected using the combined approach, resulting in a reduction in false negatives for example of 56% at 80% specificity (FIG. 25). The combined approach led to an increase of the overall performance, with an overall AUC of 0.93 (95% CI = 0.89-0.97) ($p < 0.05$ when compared to GEMINI or DELFI alone) (FIG. 4D). For stage I patients ($n=13$) the DELFI fragmentation or GEMINI analyses alone achieved AUCs of 0.73 (95% CI = 0.59-0.88) and 0.80 (95% CI = 0.67-0.93), respectively, and the combined approach resulted in an AUC of 0.87 (95% CI = 0.76-0.98) ($p < 0.05$ when compared to DELFI or GEMINI alone) (Fig. 4f). The performance of the combined GEMINI and fragmentation approach provided an overall sensitivity of 91% at a specificity of 80% (GEMINI / DELFI score > 0.38) (Table 1). When considering this approach as a pre-screen to LDCT, the sensitivity of the combined approaches with LDCT would be $> 95\%$ at a combined specificity of 85% (Table 1). Importantly, individuals with lower GEMINI / DELFI scores had better prognosis compared to individuals with higher scores ($p=0.004$, Log-rank test) (FIG. 26), reducing the concern of false negatives with this

approach, as individuals with lower scores would have better prognosis and could be detected in subsequent screens.

[0190] To externally validate the individual GEMINI method as well as the combined GEMINI / DELFI approach, an additional cohort of individuals from lung cancer screening programs were evaluated (n=57, Supplementary Table 3). This cohort included asymptomatic high-risk individuals with predominantly early stage cancers (stage I=32, II=4, III=3, IV=2 and unknown=1) where samples were collected prior to clinical diagnosis as well as individuals ultimately determined not to have cancer (n=15). Twenty one of 42 individuals with lung cancer (50%) were diagnosed with stage IA disease, similar to the proportion detected by LDCT in the National Lung Screening Trial⁵. cfDNA was isolated from plasma of these individuals and performed low coverage whole genome sequencing with coverage and feature metrics similar to the LUCAS cohort (FIGS. 27A-27D). These samples were analyzed using the fixed GEMINI and fragmentation machine learning models from the LUCAS cohort analyses. Consistent with the initial studies, it was observed that GEMINI scores were higher in high-risk individuals (age 50-80 with smoking history) with cancer compared to those without cancer (p=0.001, Wilcoxon rank sum test) (FIG. 4C). Across the validation and LUCAS cohorts, GEMINI scores of later stage lung cancer patients (stage III/IV, median GEMINI score = 0.74) were significantly higher than those of early stage patients (stage I/II, median GEMINI score = 0.64)(p=0.03, Wilcoxon rank sum test). The performance of the GEMINI approach for detecting stage I disease among individuals in this cohort was high, with an overall AUC of 0.81 (95% CI=0.67-0.94) and 0.86 (95% CI=0.74-0.97) when combined with fragmentation features (FIG. 4G). Overall, these analyses suggest that genome-wide mutational profiling is broadly generalizable for detection of early stage lung cancer in high-risk populations.

[0191] As somatic changes in lung cancer are related to smoking, it was hypothesized that there would be a relationship between the cfDNA mutational profiles and smoking history. Although overall cfDNA C>A mutation frequencies were similar among nonsmokers with and without lung cancer in the LUCAS cohort (p=0.65, Wilcoxon rank sum test), smokers with lung cancer had higher overall mutation frequencies than smokers without cancer (p=0.01, Wilcoxon rank sum test) and dramatically higher GEMINI scores (p<0.0001, Wilcoxon rank sum test) (FIGS. 23A, 23B). The GEMINI score was positively associated with years of smoking among

cancer patients ($\rho = 0.24$, $p = 0.01$). Interestingly, in individuals without cancer, the GEMINI score was negatively correlated with smoking exposure ($\rho = -0.25$, $p = 0.002$), potentially reflecting smoking-related DNA damage in non-cancer tissues³⁷ that may contribute to alterations of cfDNA. Analyses of patients in the LUCAS and validation cohorts suggested that the GEMINI approach may have higher performance in detecting individuals with greater smoking history (FIGS. 4E, 4H, 4I, 28A-28C), including an increase in GEMINI performance in the LUCAS cohort to an AUC of 0.90 and 0.95 with the combined GEMINI / DELFI approach ($p < 0.05$, DeLong's test compared to GEMINI or DELFI alone which had AUCs of 0.90 and 0.88 respectively). A positive GEMINI test at a specificity of 80% was associated with a 13.5 fold increase in the odds of cancer among ≥ 20 pack year smokers (95% CI for odds ratio: 6.7-30.7, $p < 0.0001$), and with a 20.1-fold increase among ≥ 40 pack-year smokers (95% CI for odds ratio: 7.7-54.6, $p < 0.0001$). These observations are consistent with the notion that smoking exposure results in sequence alterations in both ctDNA and non-tumor cfDNA, affecting distinct genomic regions that may be helpful for improved cancer detection using the GEMINI approach.

[0192] Given the important differences between biologic features and clinical management of SCLC and non-small cell lung cancer (NSCLC), we examined whether genome-wide mutational profiles could be used to detect SCLC and to non-invasively distinguish this cancer from other cancer types. The GEMINI score was extremely high in SCLC patients ($n=13$) compared to non-cancer individuals ($n=88$) ($p < 0.0001$, Wilcoxon rank sum test) (FIG. 5A, Supplementary Tables 2, 3), and could distinguish among these with an AUC of > 0.99 (95% CI=0.99-1.00) (FIG. 5C). The GEMINI approach was used to assess regional mutation differences in cfDNA of patients with SCLC compared to those with NSCLC ($n=99$) and found that mutation frequencies obtained in this way were higher in SCLC ($p < 0.0001$, Wilcoxon rank sum test) (FIG. 5B, Supplementary Table 4), and could be used distinguish this cancer type from NSCLC (AUC = 0.86, 95 CI=0.75-0.96) (FIG. 5C). These findings suggest that genome-wide mutation profiles may be useful for providing a non-invasive approach for detecting SCLC and distinguishing lung cancers with different histologic subtypes.

[0193] To explore the generalizability of the GEMINI approach to detect other cancers, the method was applied to evaluate a prospective cohort of individuals with or without liver cancer ($n=62$). Cross-validated regional differences in mutation frequencies identified a

significant difference in genome-wide T>C mutation profiles (FIG. 5D) in individuals with liver cancer. The derived GEMINI scores were higher in individuals with liver cancer across all stages (0-A, B, and C) compared to those with cirrhosis ($p < 0.01$ for each comparison) (FIG. 5E). Similar to analyses of lung cancer patients, GEMINI scores from patients with liver cancer were generally related to ctDNA levels, increasing with the tumor fraction estimated by ichorCNA³⁵ ($p = 0.008$, Wilcoxon rank sum test) (FIG. 23B, Supplementary Table 8).

[0194] As cfDNA mutation profiles appeared cancer-type specific, we hypothesized that the GEMINI approach could be useful for distinguishing among different cancer types. Using GEMINI, differences in mutation profiles in cfDNA were compared between NSCLC, SCLC, and HCC ($n=159$) and it was found that profiles largely clustered into three groups with each cancer type comprising the majority of observations in a cluster (FIG. 5F) (Methods). Exclusion of the most common tumor-specific alterations (FIGS. 3A-3B, 7) prevented accurate grouping by cancer type (FIG. 29). Overall, these analyses suggested that mutation profiles may be a useful method for non-invasive determination of cancer origin.

[0195] To explore whether the GEMINI approach could be used to monitor patients during therapy, we assessed serial blood samples from lung cancer patients undergoing treatment with EGFR or ERBB2 inhibitors with mutant allele fractions (MAFs) as low as 0.1%. Using the fixed model trained on the high-risk LUCAS cohort, it was found that after the initiation of therapy GEMINI scores decreased in all patients, consistent with an initial response to therapy, and over time GEMINI scores increased, consistent with the known progression of these individuals (FIG. 30). Comparison of GEMINI scores with mutant allele fractions from targeted sequencing of these patients revealed a significant correlation of ctDNA levels between the two methods (Spearman's correlation coefficient = 0.50, $p=0.03$), indicating that GEMINI has high sensitivity to low MAF levels and reflects ctDNA burden during therapy.

[0196] In this study, it was shown that individuals with cancer can be detected non-invasively through single molecule mutation profiles obtained from low coverage whole genome sequencing of cfDNA. The altered tumor-type specific mutational landscapes were detectable in plasma for cancer patients and appear to be related to replication timing and other chromatin features of the genome where repair of DNA damage may be impaired³⁸. The method described here does not require deep sequencing of matched blood cells to filter hematopoietic

alterations¹⁶, or tumor sequencing to identify tumor-specific mutations to evaluate in the plasma²², and therefore the approach is amenable for *de novo* detection and characterization of cancer. The combination of genome-wide sequence and fragmentation analyses of cfDNA provide an opportunity for cost-effective and scalable detection of cancer.

[0197] Although the majority of patients in the cohorts described represented individuals at risk for developing cancer, large-scale validation in a screening population for lung, liver and other cancers will be needed before clinical use. However, increasing the read length from 100 bp to 150 bp would increase the evaluable bases sequenced by both reads ~4 fold. Although a variety of genome-wide tumor-specific mutation profiles were assessed, including for different lung cancer histologies, liver cancer, melanoma, and lymphoma, analysis of additional genome-wide mutational profiles using other sequence alterations will likely be more effective in other settings. As mutation rates vary substantially across the cancer genome³¹, the detection of altered regional mutational frequencies in cfDNA provides generalizable approach that may be useful for early cancer detection and monitoring.

[0198] Example 2: Methods and Materials

[0199] *Study Populations Analyzed*

[0200] Tissue samples from the PCAWG Consortium consisted of 2,778 tumors with somatic mutation calls³⁹. Hypermutated tumors, including those with putative polymerase epsilon or mismatch repair defects, as well as one tumor with temozolomide treatment were excluded from analysis (n=49), as well as cancer types with less than 20 samples (n=129 samples) and cancer types with an average of <250 mutations per sample (pilocytic astrocytoma, n=89 samples) resulting in 2,511 tumors across 25 common cancer types. Single molecule mutation analyses consisted of lung cancer and matched solid tissue or blood cells from 86 donors of who passed quality control metrics³⁹. This cohort consisted of 30 females and 56 males who were diagnosed with lung cancer between ages 41 and 83. Among these individuals, 38 had lung adenocarcinoma and 48 had lung squamous cell carcinoma, and 65 of them had mutations attributed to smoking related Signature 4. Of these 65 patients, 31 of them had both tumor tissue and blood derived normal sequencing data available. Additional information regarding these samples is available in at dcc.icgc.org/releases/PCAWG. See, also Supplementary Table 1.

[0201] As previously described¹⁸, the LUCAS cohort was a prospectively collected group of 365 patients that presented in the Department of Respiratory Medicine, Infiltrate Unite, Bispebjerg Hospital, Copenhagen with a positive imaging finding on a chest X-ray or a chest CT. Patients diagnosed with cancer with known active disease or who were under treatment at the time of enrollment were excluded. The study was conducted over 7 months from September 2012 to March 2013, and all patients had a clinical follow-up until death or April 2020. All patients provided written informed consent and the studies were performed according to the Declaration of Helsinki. The LUCAS study was approved by the Danish Regional Ethics Committee and the Danish Data Protection Agency. All patients had blood samples collected at their first clinic visit before the possible diagnosis of lung cancer was made. The analyzed cohort included 158 patients with no prior, baseline, or future cancers, 114 patients with baseline lung cancer, 15 patients with a lung metastasis, and 78 patients without lung cancer at the time of blood collection, but with either earlier or later lung cancers or another cancer type. The high-risk LUCAS cohort was defined as individuals at high risk for lung cancer (age 50-80, ≥ 20 pack year smoking history) and included individuals with primary lung cancer at baseline (n=89) and individuals without prior, baseline, or future cancer (n=74). (Supplementary Table 2). There was a median ~1,451 (range: 392-2,111) haploid genome equivalents per ml of plasma analyzed.

[0202] The validation cohort was comprised of individuals from lung cancer screening programs (n=57) (Supplementary Table 3) including asymptomatic high-risk individuals with predominately early stage cancers or nodules determined to be benign. Individuals were enrolled either through the Detection of Early Lung Cancer Among Military Personnel (DECAMP) Consortium⁴⁰, or through screening efforts at the Allegheny Health Network (AHN). The DECAMP-1 protocol included current or former cigarette smokers with ≥ 20 pack-year exposure and radiological findings indicating an indeterminate pulmonary nodule of 0.7 to 3.0cm in size identified within 12 months prior to enrollment with an additional CT scan within 3 months prior to enrollment. Individuals enrolled at the AHN were identified based on eligibility for high-risk screening for lung cancer using low-dose helical CT scanning or an indication for lung cancer screening based on other high-risk characteristics such as family history of lung cancer. All patients provided written informed consent to participate in these collections and the studies were performed according to the Declaration of Helsinki. All individuals had a liquid biopsy collected prior to a possible diagnosis of lung cancer.

[0203] The lung cancer monitoring cohort consisted of serial blood draws from a cohort of lung cancer patients undergoing treatment with EGFR or ERBB2 inhibitors¹¹. The study population included samples from serial blood draws (n=18) from patients with a smoking history (n=5) with both targeted and whole-genome sequencing available¹³. Patients were age 50-73, had stage II-IV lung adenocarcinoma (n=4) or mixed histology (n=1).

[0204] The liver cancer cohort consisted of 62 patients with either liver cancer (n=48) or cirrhosis (n=14). Samples were collected prospectively as part of the HCC biomarker registry at the Johns Hopkins University School of Medicine, under a protocol approved by the Johns Hopkins Institutional Review Board. Liver cancer was defined by appropriate imaging characteristics as defined by accepted guidelines. Tumor staging was determined by the Barcelona Clinic Liver Cancer staging system (BCLC). Detailed clinical data were extracted from electronic medical records (Supplementary Table 8).

[0205] *Blood Sample Collection and Preservation*

[0206] The sample collection for the LUCAS cohort was performed at the time of the screening visit and executed as follows: venous peripheral blood was collected in one K2-EDTA tube. Within two hours from blood collection tubes were centrifuged at 2330 g at 4°C for 10 minutes. After centrifugation, EDTA plasma were aliquoted and stored at -80°C.

[0207] For the validation cohort, venous peripheral blood for each individual was collected in one K2-EDTA tube (AHN) or one Streck tube (DECAMP). Tubes from the AHN and the DECAMP collections were centrifuged at low speed (800-1600 g) for 10 minutes. The plasma portion from the first spin was spun a second time for 10 minutes. After centrifugation plasma was aliquoted and stored at -80°C for cfDNA analyses.

[0208] For the lung cancer monitoring cohort, whole blood was collected in EDTA tubes and processed immediately or within one day after storage at 4 °C or was collected in Streck tubes and processed within two days of collection as previously described¹³. Plasma and cellular components were separated by centrifugation at 800g for 10 min at 4 °C. Plasma was centrifuged a second time at 18,000g at room temperature to remove any remaining cellular debris and stored at -80 °C until the time of DNA extraction.

[0209] For the liver cancer cohort, the sample collection was performed as follows: venous peripheral blood was collected in one K2-EDTA tube. Within two hours from blood collection tubes were centrifuged at 2330 g at 4°C for 10 min, plasma was transferred to new tubes and the samples were spun at 14,000 rpm (18,000 rcf) for 10 minutes at room temperature to pellet any remaining cellular debris. After centrifugation EDTA plasma was aliquoted and stored at -80 °C for cfDNA analyses.

[0210] *Plasma Sequencing Library Preparation*

[0211] For all plasma samples, circulating cell-free DNA was isolated from 2-4 ml of plasma using the Qiagen QIAamp Circulating Nucleic Acids Kit (Qiagen GmbH), eluted in 52 µl of RNase-free water containing 0.04% sodium azide (Qiagen GmbH), and stored in LoBind tubes (Eppendorf AG) at -20°C. Concentration and quality of cfDNA were assessed using the Bioanalyzer 2100 (Agilent Technologies).

[0212] Next-generation sequencing (NGS) cfDNA libraries from the LUCAS, validation, and liver cancer cohorts were prepared for whole genome sequencing using 15 ng of cfDNA when available, or the entire purified amount when less than 15 ng was available. In brief, genomic libraries were prepared using the NEBNext DNA Library Prep Kit for Illumina (New England Biolabs (NEB)) with four main modifications to the manufacturer's guidelines: (i) the library purification steps use the on-bead AMPure XP (Beckman Coulter) approach to minimize sample loss during elution and tube transfer steps; (ii) NEBNext End Repair, A-tailing and adaptor ligation enzyme and buffer volumes were adjusted as appropriate to accommodate on-bead AMPure XP purification; (iii) Illumina dual index adaptors were used in the ligation reaction; and (iv) cfDNA libraries were amplified with Phusion Hot Start Polymerase. All of these samples underwent a 4 cycle PCR amplification after the DNA ligation step. For the lung cancer monitoring cohort, next-generation sequencing (NGS) cfDNA libraries were prepared for WGS and targeted sequencing using 5–250 ng of cfDNA as previously described^{11,13}.

[0213] *Whole Genome Sequencing Data from PCAWG Samples*

[0214] Somatic mutation calls, tumor purity, coverage statistics, as well as mutation signature abundances generated by SigProfiler²⁶ were downloaded from the International Cancer Genome Consortium (ICGC) Data Portal (<https://dcc.icgc.org/releases/PCAWG>). Bam files and

germline variant calls were downloaded from the Bionimbus Protected Data Cloud (bionimbus.opensciencedatacloud.org). Bam files were indexed using SAMtools⁴¹.

[0215] *Downsampling and Dilution of Somatic Mutations from PCAWG Lung Cancer Samples*

[0216] The downsampling and dilution experiment methodology is shown FIG. 8. Specifically, somatic mutation calls (n=3,393,564 mutations) were obtained for individuals in PCAWG with a lung cancer with the presence of Signature 4 (n=65)²⁶. Mutations were excluded with a missing value for either the number of reference or mutant alleles observed (n=5,857) resulting in 3,387,707 somatic mutations across 65 individuals. For a given individual, each observation of the reference or mutant allele separately was considered. The number of sequenced observations were computed that were tumor derived as the total number of observations multiplied by the tumor purity of the sample. We then spiked in observations with the reference allele until 10^{-1} , 10^{-2} , 10^{-3} , or 10^{-4} of the observations were of tumor origin. The average coverage of mutated positions following dilution was next computed, and the observations were randomly sampled to achieve a desired coverage of 8x, 4x, 2x, 1x, and 0.5x. For each known somatic mutation in an individual's cancer genome, we tallied the number of times that we observed the mutation for each combination of dilution amount and genome coverage and used this information to compute the percent of mutations observed in single DNA molecules.

[0217] *Whole Genome Sequencing of Plasma Samples*

[0218] Libraries prepared from whole genomes of cancer patients and cancer-free individuals were sequenced at ~2x coverage per sample using 100 bp paired-end runs (200 cycles) on the Illumina HiSeq 2000/2500 (LUCAS¹⁸, validation, and lung cancer monitoring cohorts¹³) and the NovaSeq 6000 (liver cancer cohort). To assess concordance between tissue and cfDNA mutation profiles in cancer types with few available samples, LUCAS samples were re-sequenced from patients with melanoma (n=2) and lymphoma (n=1) as well as 40 non-cancer controls and 15 individuals with largely advanced lung cancers to a median of 10x coverage on the Illumina NovaSeq 6000. Prior to alignment, adapter sequences were filtered from reads using fastp⁴². Sequence reads were aligned against the hg19 human reference genome using Bowtie2⁴³ and duplicate reads were removed using Sambamba⁴⁴. Sequencing data from

each sample comprised >7.5 million fragments, >15 million reads, >10 million reads mapped to the reference genome, >85% of bases with a Phred quality score ≥ 20 (Q20), and >80% of bases with a Phred quality score ≥ 30 (Q30).

[0219] *Identification of Single and Doublet Base Changes in Single Molecules*

[0220] The primary alignment of properly paired read pairs that mapped to autosomes in non-overlapping 100 kb bins were scanned and the base call, Phred score, and mapping quality (MAPQ) of each sequenced base using pysam was obtained. Only read pairs were considered with a MAPQ of at least 40 and only positions within each read with a Phred score of at least 30. To avoid counting multi-nucleotide variants in analyses of single base changes, positions were filtered where the two adjacent positions both contained the reference allele and had a Phred score of at least 30. A similar filter was used in analyses of doublet base changes to avoid counting larger multi-base variants. In addition, positions were removed that overlapped the Duke Excluded Regions track (hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeMapability). In each 100 kb bin, the number of sequenced bases that were C:G or A:T in the reference genome were counted. Also counted were the number of times each type of single base change (C:G>A:T, C:G>G:C, C:G>T:A, T:A>A:T, T:A>C:G, and T:A>G:C) and CC:GG>AA:TT doublet base changes in 100 kb bins were observed. Observations were counted separately based on whether the purine or the pyrimidine of each base pair was in read 1 or read 2 of the paired-end sequencing data. To exclude potential germline variants, the gnomAD database (version 3.0) was used which contains genetic variants from >70,000 whole genomes⁴⁵. The gnomAD version 3.0 variant call format (VCF) file was downloaded that was available in hg38 coordinates from the gnomAD browser. First lifted was the position of each sequence change that was identified over from hg19 to hg38 using the R package rtracklayer. Sequence changes that did not lift over to hg38, that lifted over to hg38 but to multiple different locations, or that lifted over to hg38 but the reference genome sequence differed between the hg19 and hg38 genome builds were removed. Sequence changes that were identified with their population allele frequency as well as whether the variant passed gnomAD quality filters were annotated. Any candidate variants were subsequently removed if the variant was present in gnomAD but the variant did not pass gnomAD quality filters, or if the variant was present in gnomAD with an allele frequency >1/100,000. For PCAWG samples, the

remaining variants were annotated in each sample indicating if they were called as a somatic or germline variant by the PCAWG consortium. For analyses of tissue samples, if any position in a fragment was sequenced by both read pairs, the position was kept from either read 1 or read 2 at random. For plasma samples, positions in fragments that were sequenced by both read 1 and read 2 of the read pair with the same base call were analyzed. To filter 8-oxo-dG-related sequence changes from single base analyses, any base where guanine or G>T was on read 1 and cytosine or C>A was on read 2 was excluded. To filter artifactual CC>AA changes, any bases where CC or CC>AA were on read 1 and GG or GG>TT were on read 2 were excluded. To account for potential differences in sequencing depth between samples, single molecule mutation frequencies were always computed as the number of each sequence change divided by the number of evaluable bases, defined as the number of positions in fragments in which each sequence change could be detected after quality and germline filtering.

[0221] *Estimation of 8-oxo-dG Level*

[0222] The 8-oxo-dG level was estimated for each sample as ratio of single molecule C>A frequencies when guanine or G>T was on read 1 and cytosine or C>A was on read 2 to when cytosine or C>A was on read 1 and guanine or G>T was on read 2.

[0223] *Generation of Regional Differences in Single Molecule Mutation Frequencies*

[0224] The approach to compute the regional difference in single molecule mutation frequency for a given mutation type is shown in FIG. 15. Specifically, the 100 kb bins were first aggregated to 1144 non-overlapping 2.5 Mb bins. Let y_i^0 and y_i^1 denote the number of sequence changes (e.g. C>A) at bin i for a non-cancer participant and a cancer participant, respectively. We denote the corresponding number of evaluable positions (e.g. number of C:G bases that pass quality filters) by x_i^0 and x_i^1 . The difference in the number of sequence changes at bin i relative to the number of evaluable bases comparing cancer participants to non-cancer participants for a training set comprised of $n - 1$ samples with J cancer participants and K non-cancer participants ($J + K = n - 1$) is given by

$$\delta_i = \frac{\sum_j y_{ij}^1}{\sum_j x_{ij}^1} - \frac{\sum_k y_{ik}^0}{\sum_k x_{ik}^0} \text{ for } i = 1, \dots, 1144.$$

[0225] Let $\delta_{(s)}$ denote the s^{th} order statistic such that $\delta_{(1)}$ is the bin most depleted for sequence changes in cancers relative to non-cancers and $\delta_{(1144)}$ the bin most enriched for sequence changes in cancers relative to non-cancers. Feature selection in the training set proceeds by identifying the bins at the bottom decile of δ (bins with values $\delta_{(1)}, \dots, \delta_{(114)}$) and the bins at the top decile (bins with values $\delta_{(1144)}, \dots, \delta_{(1030)}$). Denoting the bin sets for the top and bottom deciles by $\{A_{-h}\}$ and $\{B_{-h}\}$, respectively, for a training set that excludes the h^{th} sample, the regional difference in single molecule mutation frequency for the test sample is given by

$$\text{regional difference}_h = \frac{\sum_{b \in \{B_{-h}\}} y_{bh}}{\sum_{b \in \{B_{-h}\}} x_{bh}} - \frac{\sum_{a \in \{A_{-h}\}} y_{ah}}{\sum_{a \in \{A_{-h}\}} x_{ah}}$$

[0226] Using leave-one-out cross-validation, this procedure was repeated such that every participant appears in the test set once and the regional differences in single molecule mutation frequency is obtained for all n participants.

[0227] *Downsampling the Regional Difference in Single Molecule C>A Frequency to 1x Coverage in PCAWG*

[0228] For brevity, the alternative notation was used for the regional difference $\frac{y_{Bh}}{x_{Bh}} - \frac{y_{Ah}}{x_{Ah}}$ where

[0229] $y_{Bh} = \sum_{b \in \{B_{-h}\}} y_{bh}$. Denoting the down-sampled (*) regional difference by $\text{regional difference}_h^* = \frac{y_{Bh}^*}{x_{Bh}^*} - \frac{y_{Ah}^*}{x_{Ah}^*}$, these quantities were derived first by determining the number of evaluable C:G positions in the hg19 reference genome, r_A and r_B . Next, we randomly sampled (without replacement) r_A indices from the set $\{1, \dots, x_{Ah}\}$ and r_B indices from the set $\{1, \dots, x_{Bh}\}$ to represent indices of evaluable positions in these bin sets. The number of indices in the two random samples that were less than or equal to y_{Ah} and y_{Bh} were used for y_{Ah}^* and y_{Bh}^* , respectively. The above procedure was repeated until all participants in the PCAWG had a down-sampled regional difference in the single molecule C>A frequency.

[0230] *Association of the Single Molecule Mutation Frequencies with Tissue-Specific Genomic Features*

[0231] Replication timing tracks generated by the UW ENCODE group computed by averaging the wavelet-smoothed transform of the six fraction profile representing different time points during replication in 1 kb bins were downloaded from the UCSC Genome Browser from IMR90, NHEK, and GM12878 cell lines. The weighted average was computed in each 2.5 Mb bin with higher values indicating earlier replication timing. Gene expression values were obtained as raw counts using recount3⁴⁶ and converted to transcripts per million (TPM) from lung adenocarcinoma ($n = 542$), lung squamous cell carcinoma ($n = 504$), melanoma ($n = 472$), and N-cell non-Hodgkin lymphoma ($n = 48$) generated by The Cancer Genome Atlas (TCGA). For each cancer type TPM values were first averaged for each gene across samples. The gene expression in each 2.5 Mb bin in each cancer type was computed as the sum of the TPM overlapping each bin weighted by the length of the transcript. These values were then averaged between lung adenocarcinoma and lung squamous cell carcinoma to obtain a single lung cancer gene expression estimate in each bin. A/B compartmentalization data generated at 100 kb resolution through eigenvector analysis of 450K methylation array data was obtained for 12 cancer types and through eigenvector analysis of Hi-C data for GM12878 cells³³. The weighted average of the eigenvectors in 100 kb bins were computed for each 2.5 Mb bin. The average of these values from lung adenocarcinoma and lung squamous cell carcinoma was used for lung cancer analyses, GM12878 was used for BNHL analyses, and the average across all 12 cancer types was used for melanoma analyses in absence of skin A/B compartmentalization data. ChIP-seq data for H3K9me3 of A549 cells (3 pooled replicates), GM23248 cells, and Karpas 422 cells (two pooled replicates) represented as the fold change of coverage in enriched samples with respect to control samples was downloaded from the ENCODE portal (accessions: ENCFF425LVX, ENCFF098PML, and ENCFF574RYG). The weighted average of the fold changes was computed in each 2.5 Mb bin for each cell type. GC content in each 2.5 Mb bin was obtained from the hg19 reference genome. Mappability, reflecting how uniquely 100-mer sequences align to a region of the genome, was downloaded (hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeMapability/wgEncodeCrgMapabilityAlign100mer.bigWig) and aggregated into 2.5 Mb bins as the weighted average of mappability scores overlapping each bin. Genome-wide copy number was estimated for each sample using ichorCNA. Average copy number per genomic bin was computed as the weighted average of the copy number in segments overlapping each bin.

[0232] *Generation of GEMINI Scores*

[0233] To provide a calibrated score that captures the relationship between the regional difference in single molecule C>A frequency and the probability an individual has lung cancer in the high-risk LUCAS cohort, a logistic regression model was fitted for cancer status (lung GEMINI model) using the regional difference in single molecule C>A frequency as a covariate and extracted the fitted probability of cancer for each individual (lung GEMINI score). A lung GEMINI score >0.55 reflects a positive test for detection of lung cancer at 80% specificity. In addition, lung GEMINI scores were generated for the validation cohort, the cohort of patients with a baseline negative test that later developed lung cancer, the cohort of lung cancer patients that were monitored during therapy, as well as the remaining samples in the LUCAS cohort using the fixed bin sets and lung GEMINI model. For the liver cancer cohort, GEMINI scores were generated by fitting a logistic regression model for cancer status (liver GEMINI model) using the regional difference in single molecule T>C frequency as the covariate and extracting the fitted probability of cancer for each individual (liver GEMINI score). A liver GEMINI score >0.86 reflects a positive test for detection of liver cancer at 80% specificity (*See*, Supplementary Tables 1-8).

[0234] *Generation of the DELFI and Combined GEMINI / DELFI Scores*

[0235] To evaluate whether fragmentation features could further improve prediction of cancer status by GEMINI, the approach previously described¹⁸ was used on the same training sets used to generate cross-validated GEMINI scores. Briefly, the hg19 reference genome was tiled into non-overlapping 5 Mb bins. Bins with an average GC content <0.3 and an average mappability <0.9 were excluded, leaving 473 bins spanning approximately 2.4 GB of the genome. Fragment size analyses were conducted on fragments with a MAPQ of at least 30. Ratios of the number of short (100–150 bp) to long (151–220 bp) fragments across the 473 bins were normalized for GC-content and library size as previously described¹⁸. For each training set, a principal component analysis was performed on the fragmentation profiles and retained the minimum number of principal components needed to explain 90% of the variance between participants. Chromosomal arm copy number was summarized by computing a z-score for each arm using an expected coverage and standard deviation computed from an external reference set of 54 non-cancer controls (github.com/cancer-genomics/PlasmaToolsHiseq, hg19). The 39 z-

scores and principal components were integrated as covariates in a logistic regression model with a LASSO penalty. To generate DELFI scores in the validation cohort, we used the model described previously¹⁸ that was trained on 158 non-cancers and 129 cancers. The combined GEMINI / DELFI scores was computed by averaging the individual GEMINI and DELFI scores for each patient.

[0236] *Association of GEMINI Scores with the Fraction of Tumor DNA in Plasma*

[0237] The percent of tumor DNA in plasma was estimated for samples in the LUCAS and liver cancer cohorts using ichorCNA³⁵.

[0238] *Generation of the Regional Differences in Single Molecule C>A Frequencies between SCLC and NSCLC*

[0239] The regional differences in single molecule C>A frequencies were computed as previously described where individuals with SCLC were compared with those with NSCLC. To maximize the number of samples used for identifying bin sets *A* and *B*, we combined samples from the high-risk LUCAS cohort (n=10 SCLC, n=75 NSCLC) with individuals who were smokers aged 50-80 from the validation cohort (n=3 SCLC, n=24 NSCLC).

[0240] *Analysis of Different Tumor Types*

[0241] The regional difference in single molecule mutation frequency was computed as previously described by iteratively holding out each individual with either NSCLC, SCLC, or HCC (n=159) and identifying bin sets *A* and *B* using all other individuals. For each mutation type (C>A, C>G, C>T, T>A, T>C, and T>G), individuals with NSCLC were compared to those with SCLC, individuals with NSCLC were compared to those with HCC, and individuals with SCLC were compared to those with HCC, yielding 18 regional differences in mutation frequencies per individual. Principal coordinate analysis was performed on the similarity matrix generated from the Euclidean distance between pairwise samples using these 18 regional differences in mutation frequencies. K-means clustering was performed on the matrix of 18 regional differences in mutation frequencies with the number of clusters (*k*) set to 3. As a negative control, principal coordinate analysis was also performed on a similarity matrix generated from the Euclidean distance between pairwise samples after excluding C>A and T>C mutations that were most

frequently observed in lung and liver cancers, resulting in 12 regional differences in mutation frequencies per individual.

[0242] *Statistics*

[0243] The Wilcoxon rank sum test was used to generate p-values for two group comparisons. Correlation of continuous variables was performed using either the Pearson product-moment correlation coefficient or Spearman's rank correlation coefficient. All p-values were based on two-sided hypothesis tests. ROC curves were compared using DeLong's test. All confidence intervals for area under the ROC curve indicate a confidence level of 95% and were based on DeLong's method. Confidence intervals for coefficients in logistic regression models assume normality and were indicated at a 95% confidence level. An analysis of variance (ANOVA) was performed, and an F-test was used to assess whether the between sequencing lane variation of C>A frequencies or regional C>A frequencies were statistically significant. Analyses were performed with R ϵ 3.6.1 and Python 3.8.2. All boxplots represent the interquartile range with whiskers drawn to the highest value within the upper and lower fences (upper fence = 0.75 quantile + 1.5 $\times$ interquartile range; lower fence = 0.25 quantile - 1.5 $\times$ interquartile range). The solid middle line in the boxplot corresponds to the median value.

[0244] *Data Availability*

[0245] Computer code, software versions, and the computing environment for reproducing results from this study will be provided as a GitHub repository (github.com/cancer-genomics/gemini_wflow). Sequence data and clinical variables from the LUCAS study are available from the European Genome-Phenome Archive (EGA) under accession code EGAS00001005340.

Table 1. Sensitivity of GEMINI and DELFI followed by LDCT for lung cancer detection*

	GEMINI	GEMINI, DELFI	GEMINI, DELFI, LDCT
Overall (n=89)	76%	91%	93%
Overall blended (n=89)	69%	88%	91%
Stage			
Stage I (n=13)	62%	85%	88%
Stage II (n=7)	86%	86%	95%
Stage III (n=30)	73%	90%	95%
Stage IV (n=39)	82%	95%	93%
Histology			
Adenocarcinoma (n=45)	73%	89%	93%
Squamous (n=24)	71%	88%	91%
Small cell (n=10)	100%	100%	95%

*Sensitivities were determined at specificities of 80% for GEMINI or GEMINI, DELFI, or at a combined specificity of 85% for GEMINI, DELFI, LDCT with a GEMINI, DELFI pre-screen specificity of 62%. DELFI had overall sensitivity of 80% at 80% specificity. Based on the high-risk subset of patients analyzed from the LUCAS cohort¹⁸, LDCT in this setting had a specificity of 66% and an assumed sensitivity of 95%. Overall blended sensitivity reflects the sensitivity expected in a screening population weighted by the proportion of lung cancers detected in the NLST population at each stage⁵.

OTHER EMBODIMENTS

[0246] 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.

[0247] 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.

[0248] References

1. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *Ca Cancer J Clin* 71, 209–249 (2021).
2. World Health Organization. *Guide to Cancer Early Diagnosis*. (2017).
3. Moyer, V. A. U.S. Preventive Services Task Force. Screening for lung cancer: U.S. Preventive Services Task Force recommendation statement. *Annals of Internal Medicine* 160, 330–8 (2014).
4. Koning, H. J. de *et al.* Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *New Engl J Med* 382, 503–513 (2020).
5. National Lung Screening Trial Research Team. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *New Engl J Medicine* 365, 395–409 (2011).

6. Centers for Disease Control and Prevention, National Center for Health Statistics. Lung Cancer National Health Interview Survey. https://progressreport.cancer.gov/detection/lung_cancer (2021).
7. American Cancer Society. *American Cancer Society Guidelines for the Early Detection of Cancer*. (2022).
8. Phallen, J. *et al.* Direct detection of early-stage cancers using circulating tumor DNA. *Sci Transl Med* 9, eaan2415 (2017).
9. Bettegowda, C. *et al.* Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies. *Sci Transl Med* 6, 224ra24-224ra24 (2014).
10. Cohen, J. D. *et al.* Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* 359, 926–930 (2018).
11. Phallen, J. *et al.* Early Noninvasive Detection of Response to Targeted Therapy in Non-Small Cell Lung Cancer. *Cancer Res* 79, 1204–1213 (2019).
12. Newman, A. M. *et al.* Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol* 34, 547–555 (2016).
13. Cristiano, S. *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* 570, 385–389 (2019).
14. Shen, S. Y. *et al.* Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* 563, 579–583 (2018).
15. Chabon, J. J. *et al.* Integrating genomic features for non-invasive early lung cancer detection. *Nature* 580, 245–251 (2020).
16. Leal, A. *et al.* White blood cell and cell-free DNA analyses for detection of residual disease in gastric cancer. *Nat Commun* 11, 525 (2019).
17. Razavi, P. *et al.* High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. *Nat Med* 25, 1928–1937 (2019).
18. Mathios, D. *et al.* Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat Commun* 12, 5060 (2021).
19. Siejka-Zielińska, P. *et al.* Cell-free DNA TAPS provides multimodal information for early cancer detection. *Sci Adv* 7, eabh0534 (2021).
20. Wang, T.-L. *et al.* Prevalence of somatic alterations in the colorectal cancer cell genome. *Proc National Acad Sci* 99, 3076–3080 (2002).
21. Sjöblom, T. *et al.* The Consensus Coding Sequences of Human Breast and Colorectal Cancers. *Science* 314, 268–274 (2006).
22. Zviran, A. *et al.* Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat Med* 26, 1114–1124 (2020).
23. Leary, R. J. *et al.* Development of Personalized Tumor Biomarkers Using Massively Parallel Sequencing. *Sci Transl Med* 2, 20ra14-20ra14 (2010).

24. Wan, J. C. M. *et al.* Genome-wide mutational signatures in low-coverage whole genome sequencing of cell-free DNA. *Nat Commun* 13, 4953 (2022).
25. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* 578, 82–93 (2020).
26. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* 578, 94–101 (2020).
27. Chen, L., Liu, P., Evans Jr., T. C. & Ettwiller, L. M. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* 355, 752–756 (2017).
28. 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).
29. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* 354, 618–622 (2016).
30. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer genes. *Nature* 499, 214–218 (2013).
31. Gonzalez-Perez, A., Sabarinathan, R. & Lopez-Bigas, N. Local Determinants of the Mutational Landscape of the Human Genome. *Cell* 177, 101–114 (2019).
32. Mouliere, F. *et al.* Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci Transl Med* 10, eaat4921 (2018).
33. 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).
34. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* 129, 823–837 (2007).
35. Adalsteinsson, V. A. *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat Commun* 8, 1324 (2017).
36. Almodovar, K. *et al.* Longitudinal Cell-Free DNA Analysis in Patients with Small Cell Lung Cancer Reveals Dynamic Insights into Treatment Efficacy and Disease Relapse. *J Thorac Oncol* 13, 112–123 (2018).
37. Phillips, D. H. & Venitt, S. DNA and protein adducts in human tissues resulting from exposure to tobacco smoke. *Int J Cancer* 131, 2733–2753 (2012).
38. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* 521, 81–84 (2015).
39. Consortium, T. I. P.-C. A. of W. G. Pan-cancer analysis of whole genomes. *Nature* 578, 82–93 (2020).
40. Billatos, E. *et al.* Detection of early lung cancer among military personnel (DECAMP) consortium: study protocols. *Bmc Pulm Med* 19, 59 (2019).
41. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079 (2009).

42. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890 (2018).
43. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9, 357–359 (2012).
44. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 31, 2032–2034 (2015).
45. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443 (2020).
46. Wilks, C. *et al.* recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Biorxiv* 2021.05.21.445138 (2021) doi:10.1101/2021.05.21.445138.
47. Thurman, R. E., Day, N., Noble, W. S. & Stamatoyannopoulos, J. A. Identification of higher-order functional domains in the human ENCODE regions. *Genome Res* 17, 917–927 (2007).
48. The ENCODE Project Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* 489, 57–74 (2012).
49. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc National Acad Sci* 107, 139–144 (2010).

OTHER EMBODIMENTS

[0249] From the foregoing description, it will be apparent that variations and modifications may be made to the invention described herein to adopt it to various usages and conditions. Such embodiments are also within the scope of the following claims.

[0250] 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.

Supplementary Table 1. Summary of PCAWG cohort and genomic analyses*

Patient ID	Long cancer subtype	Matched normal specimen	Age at diagnosis	Sex	Whole genome coverage (tumor)	Whole genome coverage (normal)	Tumor purity	Total somatic point mutations	Fraction of mutations attributed to Signature 4	Down read for down-sampling and fusion experiment	Donor used for single molecule analyses	Estimate of copy number loss at level of tumor	Estimate of copy number loss at level of normal	Single molecule CNA changes after quality filtering (normal)	Single molecule CNA changes after quality filtering (tumor)	Regional difference in single molecule CNA frequency per Mb of exome (normal)	Regional difference in single molecule CNA frequency per Mb of exome (tumor)	Regional difference in single molecule CNA frequency after quality filtering (normal)	Single molecule CNA frequency after quality filtering (tumor)	Regional difference in single molecule CNA frequency per Mb of exome (normal)	Regional difference in single molecule CNA frequency after quality filtering (normal)	Regional difference in single molecule CNA frequency after quality filtering (tumor)	Regional difference in single molecule CNA frequency per Mb of exome (tumor)
P038459	Squamous cell carcinoma	Blood derived	53	Male	52.1	33.3	0.36	25,120	0.37	Yes	Yes	1.89	1.89	2,811,592	82.39	4.24	3.57	801,703	42.87	0.80	1.07		
P038461	Squamous cell carcinoma	Solid tissue	58	Male	48.5	45.1	0.84	43,005	0.52	Yes	No	-	-	-	-	-	-	-	-	-	-		
P038463	Squamous cell carcinoma	Blood derived	61	Male	57.8	51.2	0.76	45,829	0.49	Yes	Yes	1.64	1.64	1,402,746	57.88	13.46	14.43	1,406,173	54.84	6.73	7.30		
P038469	Squamous cell carcinoma	Solid tissue	62	Male	68.9	36.0	0.45	137,722	0.24	Yes	No	-	-	-	-	-	-	-	-	-	-		
P038475	Squamous cell carcinoma	Blood derived	60	Male	76.6	39.3	0.59	24,246	0.21	Yes	Yes	1.45	1.45	2,032,140	67.59	2.20	2.16	1,825,055	144.82	9.09	9.24		
P038476	Squamous cell carcinoma	Solid tissue	64	Male	51.5	57.7	0.63	36,861	0.42	Yes	No	-	-	-	-	-	-	-	-	-	-		
P038481	Squamous cell carcinoma	Blood derived	73	Male	72.7	36.2	0.40	22,409	0.39	Yes	Yes	1.54	1.54	3,350,900	113.74	0.66	1.20	465,392	51.75	1.71	2.29		
P038508	Squamous cell carcinoma	Solid tissue	70	Male	54.0	47.2	0.82	44,254	0.46	Yes	No	-	-	-	-	-	-	-	-	-	-		
P038512	Squamous cell carcinoma	Blood derived	60	Female	70.7	34.0	0.65	51,403	0.45	Yes	Yes	1.39	1.39	1,477,938	50.17	18.45	26.74	371,609	65.91	0.02	0.51		
P038540	Squamous cell carcinoma	Solid tissue	47	Male	67.0	35.5	0.45	38,937	0.31	Yes	No	-	-	-	-	-	-	-	-	-	-		
P038551	Squamous cell carcinoma	Solid tissue	58	Female	52.5	42.0	0.71	38,015	0.43	Yes	No	-	-	-	-	-	-	-	-	-	-		
P038563	Squamous cell carcinoma	Blood derived	75	Female	47.8	34.6	0.66	31,983	0.41	Yes	Yes	1.98	1.98	288,301	46.34	11.14	12.78	536,718	40.79	0.44	0.54		
P038572	Squamous cell carcinoma	Solid tissue	77	Female	69.6	35.9	0.74	15,202	0.26	Yes	No	-	-	-	-	-	-	-	-	-	-		
P038586	Squamous cell carcinoma	Solid tissue	52	Male	51.5	51.0	0.59	39,138	0.43	Yes	No	-	-	-	-	-	-	-	-	-	-		
P038599	Squamous cell carcinoma	Solid tissue	34	Female	54.8	40.9	0.21	47,444	0.55	Yes	No	-	-	-	-	-	-	-	-	-	-		
P038633	Squamous cell carcinoma	Solid tissue	77	Male	48.0	48.5	0.91	5,820	0	No	No	-	-	-	-	-	-	-	-	-	-		
P038691	Squamous cell carcinoma	Blood derived	58	Male	58.9	49.1	0.43	24,701	0.40	Yes	Yes	1.39	1.39	1,041,408	47.58	7.96	9.66	751,910	40.10	0.66	0.96		
P038710	Squamous cell carcinoma	Solid tissue	68	Male	61.0	35.0	0.62	61,112	0.54	Yes	No	-	-	-	-	-	-	-	-	-	-		
P038719	Squamous cell carcinoma	Blood derived	70	Male	63.6	40.0	0.50	45,123	0.46	Yes	Yes	1.64	1.64	1,508,195	52.48	13.45	15.54	1,019,583	40.34	0.78	1.17		
P038749	Squamous cell carcinoma	Solid tissue	75	Male	89.6	61.2	0.57	120,896	0.57	Yes	No	-	-	-	-	-	-	-	-	-	-		
P038796	Squamous cell carcinoma	Solid tissue	68	Male	53.0	54.0	0.80	53,051	0.38	Yes	No	-	-	-	-	-	-	-	-	-	-		
P038809	Squamous cell carcinoma	Blood derived	88	Male	56.6	58.0	0.86	54,735	0.44	Yes	Yes	1.44	0.87	1,898,050	71.50	24.40	24.25	1,846,560	141.75	1.80	5.82		

Supplementary Table 1. Summary of PLAWG cohort and genomic analyses⁴

Patient ID	Lung cancer subtype	Matched normal specimen	Age at diagnosis	Sex	Whole genome coverage (tumor)	Whole genome coverage (normal)	Tumor purity (normal)	Total somatic point mutations	Fraction of mutations attributed to Signature 4	Donor used for down-sampling and dilution experiment	Donor used for single molecule analyses	Estimate of 3,000-fold level of tumor	Estimate of 2,000-fold level of tumor (normal)	Single molecule CNA changes after quality filtering (tumor, per Mb of available bases)	Single molecule CNA frequency after quality filtering (normal, per Mb of available bases)	Regional difference in single molecule CNA frequency at 1x coverage (normal, per Mb of available bases)	Regional difference in single molecule CNA frequency after quality filtering (tumor, per Mb of available bases)	Single molecule CNA frequency after quality filtering (normal, per Mb of available bases)	Regional difference in single molecule CNA frequency after quality filtering (tumor, per Mb of available bases)	Regional difference in single molecule CNA frequency at 1x coverage (normal, per Mb of available bases)	
P0216439	Squamous cell carcinoma	Blood derived	53	Male	42.1	53.3	0.32	25,120	0.57	Yes	Yes	1.40	1.05	2,513,149	81.59	4.24	3.37	601,701	42.67	0.20	1.07
P0216441	Squamous cell carcinoma	Solid tissue	50	Male	48.5	45.1	0.64	43,065	0.52	Yes	No	-	-	-	-	-	-	-	-	-	
P0216443	Squamous cell carcinoma	Blood derived	61	Male	57.3	51.2	0.70	43,828	0.66	Yes	Yes	2.64	1.34	1,352,286	57.83	13.45	10.33	3,102,173	84.63	6.73	7.10
P0216446	Squamous cell carcinoma	Solid tissue	61	Male	68.9	36.0	0.43	137,812	0.34	Yes	No	-	-	-	-	-	-	-	-	-	
P0216475	Squamous cell carcinoma	Blood derived	20	Male	78.6	39.3	0.56	24,166	0.21	Yes	Yes	1.45	1.50	2,082,140	67.59	2.20	2.16	1,513,959	544.88	9.06	9.24
P0216476	Squamous cell carcinoma	Solid tissue	64	Male	54.5	57.7	0.65	36,581	0.42	Yes	No	-	-	-	-	-	-	-	-	-	
P0216482	Squamous cell carcinoma	Blood derived	73	Male	74.7	36.2	0.40	22,409	0.39	Yes	Yes	1.34	1.97	3,353,990	113.74	0.66	1.20	652,592	51.75	1.71	2.26
P0216506	Squamous cell carcinoma	Solid tissue	70	Male	54.0	47.2	0.62	44,251	0.66	Yes	No	-	-	-	-	-	-	-	-	-	
P0216512	Squamous cell carcinoma	Blood derived	40	Female	70.7	34.3	0.65	51,403	0.45	Yes	Yes	1.39	1.38	1,477,238	50.27	1.64	16.74	275,606	65.92	0.02	0.51
P0216540	Squamous cell carcinoma	Solid tissue	47	Male	67.0	53.5	0.40	56,840	0.51	Yes	No	-	-	-	-	-	-	-	-	-	
P0216553	Squamous cell carcinoma	Solid tissue	53	Female	92.5	43.0	0.71	39,015	0.43	Yes	No	-	-	-	-	-	-	-	-	-	
P0216553	Squamous cell carcinoma	Blood derived	73	Female	47.0	34.6	0.60	31,803	0.41	Yes	Yes	1.90	1.90	289,303	43.34	11.14	23.79	536,318	40.78	0.44	0.54
P0216571	Squamous cell carcinoma	Solid tissue	77	Female	69.6	35.9	0.74	11,302	0.26	Yes	No	-	-	-	-	-	-	-	-	-	
P0216596	Squamous cell carcinoma	Solid tissue	52	Male	51.3	51.0	0.59	31,132	0.49	Yes	No	-	-	-	-	-	-	-	-	-	
P0216598	Squamous cell carcinoma	Solid tissue	54	Female	54.2	40.9	0.22	47,444	0.55	Yes	No	-	-	-	-	-	-	-	-	-	
P0216613	Squamous cell carcinoma	Solid tissue	77	Male	49.0	48.5	0.62	5,609	0	No	No	-	-	-	-	-	-	-	-	-	
P0216632	Squamous cell carcinoma	Blood derived	56	Male	56.0	49.1	0.33	24,701	0.40	Yes	Yes	1.39	1.94	1,041,408	47.56	7.66	9.54	751,358	40.10	0.66	0.86
P0216710	Squamous cell carcinoma	Solid tissue	60	Male	61.0	35.2	0.62	51,112	0.54	Yes	No	-	-	-	-	-	-	-	-	-	
P0216719	Squamous cell carcinoma	Blood derived	70	Male	63.6	61.0	0.52	45,423	0.40	Yes	Yes	1.64	2.01	1,502,105	52.46	13.43	43.54	1,059,563	40.44	0.75	1.17
P0216768	Squamous cell carcinoma	Solid tissue	75	Male	89.6	51.2	0.57	120,800	0.57	Yes	No	-	-	-	-	-	-	-	-	-	
P0216793	Squamous cell carcinoma	Solid tissue	68	Male	53.0	54.8	0.60	55,051	0.36	Yes	No	-	-	-	-	-	-	-	-	-	
P0216806	Squamous cell carcinoma	Blood derived	49	Male	56.6	38.0	0.66	54,755	0.65	Yes	Yes	1.44	0.27	1,598,050	71.50	24.60	25.25	1,440,560	142.75	5.60	5.81

Supplementary Table 1. Summary of PLAG cohort and genomic analyses^a

Patient ID	Lung cancer subtype	Matched normal specimen	Age at diagnosis	Sex	Whole genome coverage (normal)	Whole genome coverage (tumor)	Tumor purity	Total somatic point mutations	Fraction of mutations attributed to signature 4	Driver used for down-sampling and dilution experiment	Driver used for single molecule analysis	Estimate of 48-occ-48-level (tumor)	Estimate of 48-occ-48-level (normal)	Single molecule C-A changes after germline filtering (tumor)	Single molecule C-A changes after germline filtering (normal)	Regional difference in single molecule C-A frequency after quality filtering (tumor, per Mb of evaluable bases)	Regional difference in single molecule C-A frequency after quality filtering (normal, per Mb of evaluable bases)	Single molecule C-A frequency after quality filtering (tumor, per Mb of evaluable bases)	Single molecule C-A frequency after quality filtering (normal, per Mb of evaluable bases)	Regional difference in single molecule C-A frequency after quality filtering (tumor, per Mb of evaluable bases)	Regional difference in single molecule C-A frequency after quality filtering (normal, per Mb of evaluable bases)	Regional difference in single molecule C-A frequency after quality filtering (tumor, per Mb of evaluable bases)	Regional difference in single molecule C-A frequency after quality filtering (normal, per Mb of evaluable bases)
P026684	Squamous cell carcinoma	Solid tissue	85	Male	55.9	24.4	0.75	29,450	0	NC	NC	-	-	-	-	-	-	-	-	-	-	-	-
P026684	Squamous cell carcinoma	Blood derived	81	Male	74.0	67.2	0.41	42,156	0.53	Y4	Y4	1.83	1.20	1,305,635	1,117,801	6.15	6.43	46.79	46.79	0.65	0.82	0.65	0.82
P026687	Squamous cell carcinoma	Blood derived	58	Male	55.6	65.2	0.32	37,992	0.45	Y4	Y4	1.95	2.12	1,244,552	1,062,387	8.44	7.43	40.95	40.95	1.42	1.77	1.42	1.77
P026671	Squamous cell carcinoma	Solid tissue	65	Male	74.8	42.3	0.35	37,019	0.58	Y4	Y4	-	-	-	-	-	-	-	-	-	-	-	-
P026676	Squamous cell carcinoma	Blood derived	51	Male	66.6	57.6	0.65	35,335	0.56	Y4	Y4	1.41	1.76	1,312,652	865,506	7.76	4.93	49.45	49.45	1.69	1.18	1.69	1.18
P026687	Squamous cell carcinoma	Solid tissue	64	Female	50.0	54.5	0.42	59,218	0.52	Y4	Y4	-	-	-	-	-	-	-	-	-	-	-	-
P026689	Squamous cell carcinoma	Blood derived	66	Male	69.6	41.7	0.45	55,879	0.51	Y4	Y4	1.54	1.12	1,715,656	613,4	10.96	10.96	36.45	36.45	1.96	1.66	1.96	1.66
P027035	Squamous cell carcinoma	Solid tissue	71	Male	65.4	51.5	0.39	46,579	0.42	Y4	Y4	-	-	-	-	-	-	-	-	-	-	-	-
P027087	Squamous cell carcinoma	Blood derived	62	Male	69.4	48.1	0.56	55,539	0.71	Y4	Y4	1.38	1.62	1,594,764	62,53	15.22	15.22	49.51	49.51	1.17	1.51	1.17	1.51
P027081	Squamous cell carcinoma	Solid tissue	76	Female	40.1	66.6	0.24	56,463	0.55	Y4	Y4	-	-	-	-	-	-	-	-	-	-	-	-
P027104	Squamous cell carcinoma	Solid tissue	74	Male	67.3	45.1	0.40	59,219	0.27	Y4	Y4	-	-	-	-	-	-	-	-	-	-	-	-
P027120	Squamous cell carcinoma	Blood derived	62	Female	79.9	55.8	0.47	45,814	0.44	Y4	Y4	1.79	1.84	1,540,560	45,10	10.95	10.95	37.40	37.40	0.93	0.87	0.93	0.87
P027118	Squamous cell carcinoma	Solid tissue	68	Male	44.1	49.2	0.70	79,892	0.89	Y4	Y4	-	-	-	-	-	-	-	-	-	-	-	-
P027202	Squamous cell carcinoma	Solid tissue	71	Female	66.7	31.5	0.55	42,002	0.58	Y4	Y4	-	-	-	-	-	-	-	-	-	-	-	-
P027219	Squamous cell carcinoma	Solid tissue	75	Male	52.4	36.5	0.55	62,375	0.43	Y4	Y4	-	-	-	-	-	-	-	-	-	-	-	-
P027332	Squamous cell carcinoma	Solid tissue	57	Male	70.8	34.6	0.44	75,860	0.55	Y4	Y4	-	-	-	-	-	-	-	-	-	-	-	-
P027348	Squamous cell carcinoma	Solid tissue	66	Male	63.1	53.0	0.53	77,644	0.41	Y4	Y4	-	-	-	-	-	-	-	-	-	-	-	-
P027334	Squamous cell carcinoma	Solid tissue	62	Male	74.1	54.6	0.82	40,405	0.41	Y4	Y4	-	-	-	-	-	-	-	-	-	-	-	-
P027463	Squamous cell carcinoma	Solid tissue	81	Female	65.0	31.1	0.64	30,422	0.17	Y4	Y4	-	-	-	-	-	-	-	-	-	-	-	-
P027464	Squamous cell carcinoma	Solid tissue	56	Male	92.2	81.0	0.47	59,216	0.44	Y4	Y4	-	-	-	-	-	-	-	-	-	-	-	-
P027582	Squamous cell carcinoma	Blood derived	74	Male	55.4	36.6	0.70	25,342	0	NC	NC	-	-	-	-	-	-	-	-	-	-	-	-
P027646	Squamous cell carcinoma	Blood derived	58	Male	64.2	50.3	0.73	30,266	0.39	Y4	Y4	1.37	2.09	2,174,659	47,04	7.66	7.72	49.90	49.90	1.15	1.51	1.15	1.51

Supplementary Table 1. Summary of PCWG cohort and genomic analyses^a

Patient ID	Lung cancer subtype	Matched normal specimen	Age at diagnosis	Sex	Whole genome coverage (normal)	Whole genome coverage (tumor)	Fraction of mutations attributed to signature 4	Errors used for down-sampling and diffusion experiment	Donor used for single molecule analysis	Estimate of coverage level (tumor)	Estimate of coverage level (normal)	Single molecule C>A changes after quality filtering (tumor)	Single molecule C>A frequency after quality filtering per Mb of evaluable bases	Regional difference in single molecule C>A frequency at 1x coverage (tumor, per Mb of evaluable bases)	Regional difference in single molecule C>A frequency after quality filtering (normal, per Mb of evaluable bases)	Single molecule C>A frequency after quality filtering (normal, per Mb of evaluable bases)	Regional difference in single molecule C>A frequency at 1x coverage (normal, per Mb of evaluable bases)	Regional difference in single molecule C>A frequency at 1x coverage (normal, per Mb of evaluable bases)
0017747	Squamous cell carcinoma	Salivary gland	54	Male	52.0	50.9	0.75	50,150	0.39	Yes	No	-	-	-	-	-	-	-
0017594	Adenocarcinoma	Salivary gland	62	Female	47.0	43.1	0.22	50,283	2	No	No	-	-	-	-	-	-	-
0013600	Adenocarcinoma	Bladder biopsy	59	Female	27.7	40.0	0.53	45,723	9	No	No	-	-	-	-	-	-	-

Supplementary Table 2. Effect of variant filters in tissue and ctDNA samples*

Variant filter	PRADs tumor samples (n=11)					Cumulative fold enrichment of somatic mutation frequency over background mutation frequency (Mut/100k)	PGADs blood derived normals		LUS-AS cancer ctDNA (n=69)		LUS-AS non-cancer ctDNA (n=74)		
	Known somatic mutations	Background sequence changes	Total single molecule sequence changes	Somatic single molecule mutation frequency (Mut/100k)			Background single molecule mutation frequency (Mut/100k)	Background single molecule sequence changes	Background single molecule mutation frequency (Mut/100k)	Single molecule mutation frequency changes (Mut/100k)	Single molecule mutation frequency (Mut/100k)	Background single molecule sequence changes	Background single molecule mutation frequency (Mut/100k)
No filter	8,776	3,931,580	3,940,356	5.2	2393.0	-	3,689,168	2355.3	2,801,880	1247.1	2,860,927	1244.6	
Phred < 30	6,741	571,727	578,468	5.3	448.9	5.3	540,259	420.8	944,455	467.7	958,188	462.6	
MAPQ < 40	6,713	524,475	531,187	5.5	427.6	5.8	483,980	389.1	896,155	224.3	403,275	222.4	
Blacklisted regions	6,649	519,687	526,336	5.4	425.1	5.8	488,178	386.6	887,596	220.1	394,657	218.3	
Germine variants	6,354	272,504	278,859	5.2	228.1	10.3	241,459	198.7	63,172	36.1	68,769	35.4	
8-10x-de related sequence changes	3,188	42,513	45,701	5.2	68.4	34.6	40,468	67.3	33,178	35.7	31,468	35.0	
Not in F and R reads (ctDNA only)	3,188	42,513	45,701	5.2	68.4	34.6	40,468	67.3	2,901	21.0	2,488	19.0	
Regional mutation frequency difference†				12.7	3.0	15.8		2.1		7.7		0	

*Sequence data from tissue samples were randomly downsampled to ~2x coverage similar to ctDNA samples. Values reflect the average across samples.

†Global mutation frequencies in ctDNA samples were scaled such that non-cancer samples have mean values of zero.

Supplementary Table 3. Summary of LUCAS cohort and genomic analyses*

Patient ID	Patient type	Histological diagnosis	Age at diagnosis	Sex	Stage	Smoking status	Pack years	Autoimmune disease (0=no, 1=yes)	COPD (0=no, 1=yes)	CRP (mg/L)	IL-6 (pg/mL)	Prior cancer type	Future cancer type	Time from blood draw to cancer diagnosis
C08111272P	Benign	Benign	46	Male	N/A	smoker, stopped over 6 months	25	0	0	30	2.9	N/A	N/A	N/A
C08111273P	no baseline cancer	no baseline cancer	51	Female	N/A	smoker	20	0	0	20	7.9	N/A	N/A	N/A
C08111274P	no lung cancer, other cancer, prior lung cancer	no baseline cancer	70	Female	N/A	never smokes	0	0	0	20	5.2	breast cancer operated	uterine cancer, breast soft tissue sarcoma, mesothelioma, renal medullary RCC	-2222, 439, 1495, 1814, 2435
C08111275P	no lung cancer, other cancer	Benign	60	Male	N/A	smoker, stopped over 6 months	40	0	0	27	7.7	N/A	colorectal cancer	1471
C08111276P	no baseline cancer	no baseline cancer	71	Male	N/A	smoker	100	0	1	30	11.9	N/A	N/A	N/A
C08111277P	benign	Benign	23	Male	N/A	smoker, stopped over 6 months	1	0	0	20	6.5	N/A	N/A	N/A
C08111278P	no baseline cancer	no baseline cancer	50	Male	N/A	smoker	43	0	0	20	4.1	N/A	N/A	N/A
C08111279P	no lung cancer, other cancer	Benign	65	Female	N/A	smoker, stopped over 6 months	25	1	0	20	26	cellular carcinoma of skin, squamous	N/A	-4877
C08111280P	no lung cancer, other cancer	Benign	71	Male	N/A	smoker	50	0	1	25	7.8	bladder cancer stage II	N/A	-3945
C08111281P	lung cancer	Squamous	60	Female	III	smoker, stopped over 6 months	3	0	0	22	8.1	N/A	N/A	N/A
C08111282P	no baseline cancer	no baseline cancer	24	Female	N/A	never smokes	0	1	0	20	4.1	N/A	N/A	N/A
C08111283P	lung cancer	adenocarcinoma	75	Male	IV	smoker, stopped over 6 months	50	0	0	30	2.3	N/A	N/A	N/A
C08111284P	lung cancer	adenocarcinoma	62	Male	IV	smoker, stopped over 6 months	80	0	1	36	566	N/A	N/A	N/A
C08111285P	no lung cancer, other cancer	no baseline cancer	60	Female	N/A	smoker, stopped over 6 months	3	0	0	30	3.2	colorectal cancer in situ, operated	N/A	-9232
C08111286P	no baseline cancer	no baseline cancer	56	Female	N/A	smoker, stopped over 6 months	0	1	0	20	1.9	N/A	N/A	N/A
C08111287P	no baseline cancer	no baseline cancer	60	Female	N/A	smoker, stopped over 6 months	13	0	0	20	5.5	N/A	N/A	N/A
C08111288P	no baseline cancer	no baseline cancer	55	Male	N/A	smoker	45	0	0	30	14.2	N/A	N/A	N/A
C08111289P	lung cancer, prior lung cancer	adenocarcinoma	55	Male	N/A	smoker	37	0	0	20	2.6	N/A	N/A	N/A
C08111290P	no lung cancer	no baseline cancer	71	Female	IA	smoker, stopped over 6 months	35	0	1	20	3.9	lung cancer metastatic adenocarcinoma	N/A	-6523
C08111291P	no baseline cancer	no baseline cancer	75	Female	N/A	smoker, stopped over 6 months	1	1	0	20	1	N/A	N/A	N/A
C08111292P	no lung cancer, other cancer	no baseline cancer	60	Female	N/A	never smokes	0	0	0	20	4.3	colorectal cancer, squamous, squamous, mesothelioma, operated	N/A	-147
C08111293P	no lung cancer, other cancer	no baseline cancer	70	Female	N/A	smoker	55	1	0	30	3.4	breast cancer operated	N/A	-10945
C08111294P	lung cancer	adenocarcinoma	78	Female	IV	smoker	30	0	0	41	15.7	N/A	N/A	N/A

Patient ID	Patient type	Histological diagnosis	Age at diagnosis	Sex	Stage	Smoking status	Pack years	Autism spectrum disease (DSM-5, 1=Yes)	COPD (DSM-5, 1=Yes)	CRP (mg/L)	PLT (mg/dL)	Prior cancer type	Future cancer type	Time from blood draw to cancer diagnosis
COP111290P	No baseline cancer	No baseline cancer	37	Male	N/A	never smoker	0	0	0	10	2.2	N/A	N/A	N/A
COP111291P	No lung cancer, other cancer	No baseline cancer	75	Male	N/A	smoker, stopped over 6 months	15	0	0	10	14.4	N/A	basal cell carcinoma	1275
COP111292P	No baseline cancer	No baseline cancer	65	Female	N/A	smoker, stopped over 6 months	40	0	0	10	1.1	N/A	N/A	N/A
COP111293P	benign	benign	60	Female	N/A	smoker, stopped over 6 months	1	1	0	10	N/A	N/A	N/A	N/A
COP111294P	No lung cancer, other cancer	benign	57	Male	N/A	smoker	60	0	0	23	30.1	N/A	basal cell carcinoma, squamous cell carcinoma	203, 1543
COP111295P	No baseline cancer	No baseline cancer	24	Female	N/A	never smoker	0	0	0	10	1.4	N/A	N/A	N/A
COP111296P	lung cancer	adenocarcinoma	60	Male	IIIB	smoker	10	0	0	10	4.4	N/A	N/A	N/A
COP111297P	lung cancer	squamous	51	Female	IIIA	smoker	35	0	1	12	0.2	N/A	N/A	N/A
COP111298P	No baseline cancer	No baseline cancer	60	Female	N/A	smoker, stopped over 6 months	62	0	0	10	9.9	N/A	N/A	N/A
COP111299P	No baseline cancer	No baseline cancer	65	Female	N/A	smoker	46	1	0	10	10.1	N/A	N/A	N/A
COP111300P	No baseline cancer	No baseline cancer	71	Male	N/A	smoker, stopped over 6 months	34	0	0	10	2.9	N/A	N/A	N/A
COP111301P	lung cancer	adenocarcinoma	50	Male	IV	smoker, stopped over 6 months	25	0	0	04	10.7	N/A	N/A	N/A
COP111302P	lung cancer	adenocarcinoma	61	Male	III	smoker	40	0	0	11	22.4	N/A	N/A	N/A
COP111303P	benign	benign	44	Male	N/A	smoker	30	0	0	10	2.9	N/A	N/A	N/A
COP111304P	benign	benign	70	Female	N/A	smoker, stopped over 6 months	20	1	1	10	4.5	N/A	N/A	N/A
COP111310P	No baseline cancer	No baseline cancer	37	Female	N/A	smoker	30	1	0	10	2.3	N/A	N/A	N/A
COP111311P	lung cancer	adenocarcinoma	20	Male	IV	smoker, stopped over 6 months	25	0	1	10	0.9	N/A	N/A	N/A
COP111312P	No lung cancer, other cancer	No baseline cancer	71	Male	N/A	smoker	17	0	0	10	2.7	N/A	squamous cell carcinoma	634
COP111313P	lung cancer	adenocarcinoma	65	Male	IV	smoker, stopped over 6 months	0	0	0	27	0.2	N/A	N/A	N/A
COP111314P	lung cancer	adenocarcinoma	72	Male	IV	smoker, stopped over 6 months	40	0	0	11	1.2	prostate cancer, operated	N/A	-922
COP111315P	No baseline cancer	No baseline cancer	54	Male	N/A	smoker	40	0	0	17	15.0	N/A	N/A	N/A
COP111316P	lung cancer	squamous	71	Male	IIIB	smoker	50	0	0	10	5.2	N/A	N/A	N/A
COP111317P	No baseline cancer	No baseline cancer	59	Male	N/A	never smoker	0	0	0	10	1.1	N/A	N/A	N/A
COP111318P	No lung cancer, other cancer	No baseline cancer	61	Male	N/A	smoker	23	0	0			Superficial spreading melanoma, excised	basal cell carcinoma	-1434, -131
COP111319P	No lung cancer, other cancer	No baseline cancer	61	Male	N/A	smoker, stopped over 6 months	20	0	0	10	11.0	prostate cancer, operated	N/A	-1770
COP111320P	lung cancer	adenocarcinoma	71	Female	IV	smoker, stopped over 6 months	40	0	1	10	3.3	N/A	basal cell carcinoma	N/A
COP111321P	lung cancer	squamous	65	Female	IIIA	unknown	20	0	0	22	11.8	N/A	N/A	N/A

Patient ID	Patient type	Histological diagnosis	Age at diagnosis	Sex	Stage	Smoking status	Post years	Autoimmune disease (y=no, 1=yes)	COPD (y=no, 1=yes)	CEP (mg/L)	HLA-DQ (mg/mL)	Prior cancer type	Future cancer type	Time from blood draw to cancer diagnosis
COP10323P	Benign	Benign	47	Female	N/A	Smoker, stopped over 6 months	3	0	0	2.0	2.1	N/A	N/A	N/A
COP10323P	lung cancer, other cancer	Squamous	62	Female	3	Smoker, stopped over 6 months	40	1	0	2.0	2.1	uterine cancer in situ, operated	N/A	-7539
COP10324P	no lung cancer, other cancer	No baseline cancer	79	Female	N/A	Smoker	20	1	0	2.0	2.9	Basal cell carcinoma	N/A	-3872
COP10323P	lung cancer	Adenocarcinoma	43	Female	3A	Smoker	40	1	0	1.1	7.1	N/A	N/A	N/A
COP10324P	no lung cancer, other cancer	No baseline cancer	52	Female	N/A	Smoker, stopped over 6 months	4	0	0	2.0	1.7	uterine cancer in situ, operated	N/A	-4297
COP10327P	lung cancer	SCLC	56	Female	3B	Smoker	40	1	1	1.0	2.2	N/A	N/A	N/A
COP10323P	no lung cancer, other cancer	No baseline cancer	32	Female	N/A	Smoker, stopped over 6 months	15	0	0	1.9	N/A	Basal cell carcinoma, squamous cancer disseminated	N/A	-340, 614, 2576
COP10323P	no lung cancer, other cancer	No baseline cancer	77	Male	N/A	Smoker, stopped over 6 months	35	0	1	1.5	2.2	Basal cell carcinoma	N/A	-1082
COP10323P	no baseline cancer	No baseline cancer	68	Male	N/A	Smoker, stopped over 6 months	10	0	0	2.0	1.7	N/A	N/A	N/A
COP10321P	lung cancer, other cancer	Adenocarcinoma	71	Female	3A	Smoker	22	0	1	2.0	1.4	N/A	Breast cancer	1246
COP10323P	Benign	Benign	71	Female	N/A	Smoker, stopped over 6 months	27	1	0	2.3	10.3	N/A	N/A	N/A
COP10323P	lung cancer	Squamous	65	Female	N/A	Smoker, stopped over 6 months	45	1	0	2.9	14.6	N/A	N/A	N/A
COP10324P	no baseline cancer	No baseline cancer	65	Male	N/A	Smoker	50	0	1	4.0	265	N/A	N/A	N/A
COP10325P	lung cancer	Squamous	61	Male	7	Smoker, stopped over 6 months	45	0	0	42	17.6	N/A	N/A	N/A
COP10323P	lung cancer	Adenocarcinoma	59	Female	3B	Smoker	55	1	0	1.0	4.2	N/A	N/A	N/A
COP10327P	no lung cancer, other cancer	No baseline cancer	61	Female	N/A	Never smoker	0	1	0	2.0	0.84	Squamous cell carcinoma, operated	N/A	-213
COP10323P	Benign	Benign	64	Female	N/A	Never smoker	0	1	0	2.0	2.4	N/A	N/A	N/A
COP10320P	Benign	Benign	29	Male	N/A	Never smoker	0	0	0	2.0	1.7	N/A	N/A	N/A
COP10341P	lung cancer	Squamous	79	Male	3B	Smoker	40	0	0	2.4	11.4	N/A	N/A	N/A
COP10342P	no lung cancer, other cancer	No baseline cancer	75	Male	N/A	Smoker	120	0	0	2.0	5.2	N/A	Prostate cancer	952
COP10343P	lung cancer, other cancer	SCLC	59	Male	3B	Smoker, stopped over 6 months	45	0	0	3.0	20.3	Bladder cancer, operated	N/A	-2326
COP10344P	lung cancer	Squamous	71	Male	3A	Smoker, stopped over 6 months	35	1	0	2.1	2.3	N/A	N/A	N/A
COP10345P	no lung cancer, other cancer	Benign	71	Male	N/A	Smoker	110	0	0	2.5	44.7	thyroid cancer, operated	N/A	-4201
COP10346P	no baseline cancer	No baseline cancer	69	Female	N/A	Smoker, stopped over 6 months	50	0	1	2.0	2	N/A	N/A	N/A
COP10347P	metastases from another primary cancer	Metastases from another primary cancer	64	Male	7	Smoker, stopped over 6 months	20	1	0	1.5	11.5	superficial spreading malignant melanoma, operated	N/A	-3827
COP10342P	Benign	Benign	67	Male	N/A	Smoker	45	0	1	3.5	97.6	N/A	N/A	N/A

Patient ID	Patient type	Histological diagnosis	Age at diagnosis	Sex	Stage	Smoking status	Pack years	Autoimmune disease (0=No, 1=Yes)	COPD (0=No, 1=Yes)	CrP (mg/dL)	IL-6 (pg/mL)	Prior cancer type	Future cancer type	Time from blood draw to cancer diagnosis
COP110340P	lung cancer, other cancer	adenocarcinoma	71	Female	N/A	smoker	40	0	1	13	9.7	breast cancer operated	N/A	2512
COP110350P	no lung cancer, other cancer	no baseline cancer	73	Female	N/A	smoker, stopped over 5 months	40	0	1	13	9.2	N/A	N/A	1654
COP110351P	lung cancer, benign	Benign	40	Male	N/A	smoker, stopped	0	1	0	13	1.02	N/A	N/A	N/A
COP110352P	no baseline cancer	no baseline cancer	41	Male	N/A	smoker, stopped over 5 months	11	0	0	13	2.3	N/A	N/A	N/A
COP110353P	benign	Benign	60	Male	N/A	smoker, stopped over 5 months	30	0	1	34	14.2	N/A	N/A	N/A
COP110354P	no baseline cancer	no baseline cancer	43	Female	N/A	smoker, stopped over 5 months	22	0	0	13	1.1	N/A	N/A	N/A
COP110355P	met to lung, other cancer	Metastasis from another primary cancer	76	Male	IV	smoker, stopped over 5 months	1	0	0	10	7.6	rectal cancer operated	rectal cancer, recurrent, pericardial cancer	-1117, 1112, 2036
COP110356P	no lung cancer, other cancer	Benign	20	Male	N/A	smoker, stopped over 5 months	10	1	0	11	6.7	squamous cell carcinoma, basal cell carcinoma operated	basal cell carcinoma	-1687, *720, 1433
COP110357P	no lung cancer, other cancer	no baseline cancer	52	Female	N/A	smoker, stopped over 5 months	12	1	0	10	4.1	renal cancer (papillary, renal cancer), sigmoidal cancer	sigmoidal cancer, recurrent	-2643, -2135, -1770, 1159
COP110358P	benign	Benign	55	Female	N/A	smoker	30	0	0	10	1.1	N/A	N/A	N/A
COP110359P	no baseline cancer	no baseline cancer	53	Female	N/A	smoker	40	0	1	10	4.9	N/A	N/A	N/A
COP110360P	benign	Benign	35	Male	N/A	smoker	35	1	0	10	19	N/A	N/A	N/A
COP110361P	no baseline cancer	no baseline cancer	29	Female	N/A	smoker, stopped over 5 months	30	1	1	10	4.1	N/A	N/A	N/A
COP110362P	no baseline cancer	no baseline cancer	71	Female	N/A	smoker, stopped over 5 months	20	0	1	13	6.5	N/A	N/A	N/A
COP110363P	no lung cancer, other cancer	no baseline cancer	22	Male	N/A	smoker	33	0	1	10	10.1	N/A	prostate cancer, melanoma, colon cancer, adenocarcinoma	121, 377, 169
COP110364P	benign	Benign	41	Male	N/A	smoker	25	0	0	11	1.5	N/A	N/A	N/A
COP110365P	no lung cancer, other cancer	no baseline cancer	52	Female	N/A	smoker, stopped over 5 months	11	0	0	10	1.6	nodular lymphoma, basal cell carcinoma	basal cell carcinoma	-3031, -1326, 797
COP110366P	lung cancer, prior lung cancer	Squamous	34	Male	IV	smoker, stopped over 5 months	70	0	0	66	41.3	lung cancer, nodular lymphoma	N/A	-803
COP110367P	no baseline cancer	no baseline cancer	79	Female	N/A	smoker, stopped over 5 months	40	0	0	13	1.2	N/A	N/A	N/A
COP110368P	no baseline cancer	no baseline cancer	21	Female	N/A	smoker, stopped over 5 months	25	0	1	15	7.9	N/A	N/A	N/A
COP110369P	lung cancer, other cancer	adenocarcinoma	55	Female	IV	unknown	35	0	0	17	10	uterine cancer operated	basal cell carcinoma	*743
COP110370P	no lung cancer, other cancer	no baseline cancer	59	Female	N/A	never smoker	0	1	0	13	1.1	N/A	breast cancer operated	130

Patient ID	Patient type	Histological diagnosis	Age at diagnosis	Sex	Stage	Smoking status	Pack years	Autoimmune disease (F-Ho, 1=Yes)	COPD (F-Ho, 1=Yes)	CRP (mg/L)	IL-6 (pg/mL)	First cancer type	Future cancer type	Time from blood draw to cancer diagnosis
CSP411371P	no lung cancer other cancer	no baseline cancer	75	Female	N/A	smoker	40	0	0	1.0	2.3	basal cell carcinoma	N/A	-1901
CSP411372P	no baseline cancer	no baseline cancer	63	Female	N/A	smoker, stopped smoker 6 months	25	0	0	1.0	2.3	N/A	N/A	N/A
CSP411373P	lung cancer, other cancer	Squamous	62	Female	N/A	smoker, stopped over 6 months	10	0	0	20	21	breast cancer, osteosarcoma	N/A	-1972
CSP411374P	benign	benign	35	Female	N/A	smoker	50	0	1	1.0	4	N/A	N/A	N/A
CSP411375P	lung cancer	adenocarcinoma	60	Male	IV	smoker, stopped over 5 months	10	0	1	353	1076	N/A	N/A	N/A
CSP411376P	lung cancer, other cancer	adenocarcinoma	70	Female	IA	smoker, stopped over 5 months	22	0	0	1.0	5.3	breast cancer, osteosarcoma, melanoma, sarcoma	adenocarcinoma sarcoma	-1987, -2019, -1999
CSP411377P	benign	benign	65	Female	N/A	smoker, stopped over 5 months	15	0	0	1.0	59	N/A	N/A	N/A
CSP411378P	lung cancer	Squamous	50	Male	IV	smoker, stopped over 5 months	24	1	0	63	23	N/A	N/A	N/A
CSP411379P	no baseline cancer	no baseline cancer	59	Male	N/A	smoker, stopped over 5 months	35	0	0	1.0	1.2	N/A	N/A	N/A
CSP411380P	benign	benign	40	Male	N/A	smoker	30	0	0	1.0	1.5	N/A	N/A	N/A
CSP411381P	no baseline cancer	no baseline cancer	46	Female	N/A	smoker, stopped over 6 months	35	0	0	1.0	0.2	N/A	N/A	N/A
CSP411382P	benign	benign	53	Male	N/A	never smoker	0	0	0	1.0	1.99	N/A	N/A	N/A
CSP411383P	met to lung cancer	metastasis from smoker primary cancer	43	Male	IV	never smoker	0	0	0	22	12.4	N/A	N/A	N/A
CSP411384P	lung cancer	adenocarcinoma	66	Female	IV	smoker	21	1	0	1.0	9.1	N/A	N/A	N/A
CSP411385P	no baseline cancer	no baseline cancer	55	Female	N/A	smoker, stopped over 5 months	12	0	0	21	0.7	N/A	N/A	N/A
CSP411386P	no baseline cancer	no baseline cancer	61	Male	N/A	never smoker	0	0	0	1.0	1.2	N/A	N/A	N/A
CSP411387P	no baseline cancer	no baseline cancer	51	Female	N/A	smoker, stopped over 6 months	25	0	0	1.0	1	N/A	N/A	N/A
CSP411388P	lung cancer	Squamous	75	Male	IV	smoker, stopped over 6 months	75	0	1	144	20.3	N/A	N/A	N/A
CSP411389P	no baseline cancer	no baseline cancer	67	Male	N/A	smoker	47	0	0	1.0	2.2	N/A	N/A	N/A
CSP411390P	lung cancer, prior lung cancer	Unknown	74	Male	IV	smoker, stopped over 6 months	50	1	0	23	72.5	lung cancer, metastatic breast	N/A	-1934
CSP411391P	benign	benign	56	Male	N/A	smoker, stopped over 6 months	5	0	0	1.0	1.6	N/A	N/A	N/A
CSP411392P	lung cancer	Squamous	70	Male	IV	unknown	N/A	0	0	26	9.7	N/A	N/A	N/A
CSP411393P	no baseline cancer	no baseline cancer	53	Female	N/A	smoker, stopped over 6 months	30	0	0	1.0	6.52	N/A	N/A	N/A
CSP411394P	no baseline cancer	no baseline cancer	62	Male	N/A	smoker, stopped over 5 months	5	0	0	32	25	N/A	N/A	N/A
CSP411395P	no baseline cancer	no baseline cancer	64	Female	N/A	never smoker	0	1	0	1.0	0.98	N/A	N/A	N/A
CSP411397P	lung cancer, other cancer	SOLC	57	Female	IB	smoker	40	0	0	1.0	1.58	breast cancer, osteosarcoma	N/A	-191
CSP411398P	lung cancer	adenocarcinoma	62	Male	IV	smoker, stopped over 6 months	40	0	0	1.0	1.31	N/A	N/A	N/A

Patient ID	Patient type	Histological diagnosis	Age at diagnosis	Sex	Stage	Smoking status	Pack years	Autoimmune disease (0-100, 1=Yes)	CRP (mg/L)	IL-6 (pg/mL)	Prior cancer type	Future cancer type	Time from blood draw to cancer diagnosis
CSP1110390	no lung cancer, other cancer	No baseline cancer	71	Male	N/A	smoker, stopped over 6 months	43	1	10	13.2	N/A	prostate cancer, bladder cancer	734, 1412
CSP1110400	lung cancer	Adenocarcinoma	66	Female	I/II	smoker, stopped over 6 months	27	0	12	16.3	N/A	N/A	N/A
CSP1110410	no baseline cancer	No baseline cancer	52	Male	N/A	smoker	30	1	20	4.9	N/A	N/A	N/A
CSP1110420	benign	Benign	46	Male	N/A	smoker	30	0	10	2.2	N/A	N/A	N/A
CSP1110430	lung cancer	Adenocarcinoma	59	Female	IIIA	smoker	23	1	11	17.8	N/A	N/A	N/A
CSP1110440	lung cancer	SCLC	73	Female	I/II	smoker	40	1	36	13.5	N/A	N/A	N/A
CSP1110450	benign	Benign	64	Female	N/A	smoker, stopped over 6 months	20	1	10	2.7	N/A	N/A	N/A
CSP1110460	no baseline cancer	No baseline cancer	57	Male	N/A	smoker, stopped over 6 months	15	1	10	1.1	N/A	N/A	N/A
CSP1110470	no baseline cancer	No baseline cancer	66	Male	N/A	smoker, stopped over 6 months	45	0	10	2.6	N/A	N/A	N/A
CSP1110480	lung cancer	Adenocarcinoma	45	Male	I/II	smoker, stopped over 6 months	45	0	11	1	N/A	N/A	N/A
CSP1110490	lung cancer	Adenocarcinoma	59	Female	I/II	never smoker	0	0	20	1.1	N/A	N/A	N/A
CSP1110500	lung cancer, other cancer	Adenocarcinoma	66	Female	IIIB	smoker	40	0	10	3.8	N/A	colon cancer	2513
CSP1110510	no lung cancer, other cancer	No baseline cancer	65	Female	N/A	smoker, stopped over 6 months	4	1	10	2.5	N/A	basal cell carcinoma	1809
CSP1110520	lung cancer	Adenocarcinoma	52	Female	I/II	smoker	30	0	17	11.9	N/A	N/A	N/A
CSP1110530	benign	Benign	56	Female	N/A	smoker	40	0	19	1.5	N/A	N/A	N/A
CSP1110540	lung cancer	Adenocarcinoma	71	Male	IIIB	smoker, stopped over 6 months	40	0	36	17.9	N/A	N/A	N/A
CSP1110550	lung cancer	Adenocarcinoma	67	Male	IIIB	smoker, stopped over 6 months	33	1	10	2.3	N/A	N/A	N/A
CSP1110560	no lung cancer, other cancer	No baseline cancer	64	Female	N/A	smoker	45	0	10	0.0	N/A	N/A	N/A
CSP1110570	lung cancer	Adenocarcinoma	64	Male	N/A	smoker, stopped over 6 months	30	0	10	2.7	prostate cancer, colorectal cancer	N/A	?
CSP1110580	lung cancer	Squamous	62	Female	II	smoker, stopped over 6 months	35	0	42	13.7	N/A	N/A	N/A
CSP1110590	benign	Benign	45	Male	N/A	never smoker	0	0	10	3.9	N/A	N/A	N/A
CSP1110600	benign	Benign	62	Male	N/A	smoker, stopped over 6 months	40	1	10	2.5	N/A	N/A	N/A
CSP1110610	no baseline cancer	No baseline cancer	65	Female	N/A	never smoker	0	1	10	2.5	N/A	N/A	N/A
CSP1110620	benign	Benign	65	Female	N/A	smoker	40	0	10	1.1	N/A	N/A	N/A
CSP1110630	lung cancer	Metastatic	75	Male	IIIB	never smoker	0	0	10	0.0	N/A	N/A	N/A
CSP1110640	lung cancer	Squamous	59	Female	IIIB	smoker, stopped over 6 months	40	0	11	10	3.7	N/A	N/A
CSP1110650	no baseline cancer	No baseline cancer	56	Female	N/A	smoker	25	0	10	1.2	N/A	N/A	N/A
CSP1110660	lung cancer	SCLC	63	Female	IIIA	smoker, stopped over 6 months	35	0	10	5.7	N/A	N/A	N/A
CSP1110670	no lung cancer, prior lung cancer	Benign	76	Female	N/A	smoker	25	0	10	4.3	lung cancer, NSCLC, operated aden	N/A	1536
CSP1110680	no baseline cancer	No baseline cancer	63	Male	N/A	smoker, stopped over 6 months	47	1	10	6	N/A	N/A	N/A

Patient ID	Patient type	Histological diagnosis	Age at diagnosis	Sex	Stage	Smoking status	Pack years	Autoimmune disease (0=no, 1=yes)	CDP (0=no, 1=yes)	CEP (ng/L)	IL-6 (pg/mL)	Prior cancer type	Future cancer type	Time from blood draw to cancer diagnosis
OSP110423P	met to lung	Metastases from another primary cancer	81	Female	IV	smoker, stopped over 6 months	1	0	0	13	0.7	N/A	N/A	N/A
OSP110430P	lung cancer	Metastases from another primary cancer	76	Female	IV	smoker, stopped over 6 months	12	0	1	16	7.6	N/A	N/A	N/A
OSP110431P	met to lung, other cancer	Metastases from another primary cancer	56	Female	IV	never smoker	0	0	0	26	9.4	breast cancer, unspecified	recurrent breast cancer	~8.18
OSP110432P	no baseline cancer	No baseline cancer	70	Male	N/A	smoker	85	0	0	1.8	4.2	N/A	N/A	N/A
OSP110433P	no lung cancer, other cancer	No baseline cancer	59	Male	N/A	smoker	10	0	0	24	1.8	basal cell carcinoma, unspecified	basal cell carcinoma, unspecified	~3813, ~3209, 925, 2386
OSP110434P	no lung cancer, other cancer	No baseline cancer	77	Male	N/A	smoker, stopped over 6 months	50	0	1	25	23.0	basal cell carcinoma	N/A	~2846
OSP110435P	lung cancer, other cancer	Adenocarcinoma	67	Female	IA	smoker, stopped over 6 months	23	0	0	10	0.93	basal cell carcinoma	N/A	~7799
OSP110436P	benign	Benign	76	Male	N/A	never smoker	0	0	0	10	1.1	N/A	N/A	N/A
OSP110437P	lung cancer	Adenocarcinoma	82	Female	IV	smoker, stopped over 6 months	50	1	1	10	6.3	N/A	N/A	N/A
OSP110438P	no lung cancer, other cancer	Benign	83	Male	N/A	smoker, stopped over 6 months	70	0	1	45	14.4	basal cell carcinoma, unspecified	N/A	~8279, ~3480, ~518
OSP110439P	lung cancer	Adenocarcinoma	81	Female	IV	smoker, stopped over 6 months	80	0	0	10	2.5	N/A	N/A	N/A
OSP110440P	no baseline cancer	No baseline cancer	52	Male	N/A	smoker	40	0	0	14	1.5	N/A	N/A	N/A
OSP110441P	lung cancer	Adenocarcinoma	64	Female	IV	smoker, stopped over 6 months	50	0	0	10	6.6	N/A	N/A	N/A
OSP110442P	no baseline cancer	No baseline cancer	82	Female	N/A	smoker	25	0	0	10	2.2	N/A	N/A	N/A
OSP110443P	lung cancer	Adenocarcinoma	73	Male	III	smoker	37	0	0	10	5.1	N/A	N/A	N/A
OSP110444P	no lung cancer, other cancer	No baseline cancer	80	Male	N/A	smoker, stopped over 6 months	35	0	0	10	6.7	N/A	multiple myeloma	~7707
OSP110445P	no baseline cancer	No baseline cancer	60	Male	N/A	smoker	50	0	1	10	1.7	N/A	N/A	N/A
OSP110446P	benign	Benign	80	Male	N/A	smoker	42	0	0	10	2.5	N/A	N/A	N/A
OSP110447P	met to lung	Metastases from another primary cancer	55	Male	IV	smoker, stopped over 6 months	45	0	0	10	1.8	N/A	N/A	N/A
OSP110448P	no lung cancer, other cancer	Benign	85	Male	N/A	smoker, stopped over 6 months	15	0	1	36	7.2	colon cancer, unspecified, adenocarcinoma	N/A	~45
OSP110449P	lung cancer	Squamous	80	Female	III	smoker	40	0	1	13	22.9	N/A	N/A	N/A
OSP110450P	no lung cancer, other cancer	No baseline cancer	74	Male	N/A	smoker, stopped over 6 months	30	0	0	10	4.1	diffuse large B-cell lymphoma	N/A	~1811
OSP110451P	benign	Benign	37	Male	N/A	smoker	11	0	0	10	3.1	N/A	N/A	N/A
OSP110452P	lung cancer	Adenocarcinoma	62	Female	IV	smoker, stopped over 6 months	29	0	0	42	7.09	N/A	N/A	N/A
OSP110453P	no baseline cancer	No baseline cancer	56	Male	N/A	smoker, stopped over 6 months	40	0	0	10	2.5	N/A	N/A	N/A
OSP110454P	benign	Benign	59	Male	N/A	smoker, stopped over 6 months	43	0	1	10	2.5	N/A	N/A	N/A
OSP110455P	no baseline cancer	No baseline cancer	71	Female	N/A	smoker, stopped over 6 months	7	0	0	10	1.9	N/A	N/A	N/A
OSP110456P	lung cancer	Squamous	52	Male	N/A	smoker	35	1	1	40	33.8	N/A	N/A	N/A

Patient ID	Patient type	Histological diagnosis	Age at diagnosis	Sex	Stage	Smoking status	Pack years	Autoimmune disease (0=no, 1=yes)	CEPT (0=no, 1=yes)	CRP (mg/L)	IL-6 (pg/mL)	Prior cancer type	Future cancer type	Time from blood draw to cancer diagnosis
CER1114579	lung cancer	NSCL	70	Female	IIIb	smoker, stopped over 5 months	30	0	0	12	11.9	NS	NS	NS
CER1114580	lung cancer	NSCL	59	Male	IIIc	smoker, stopped over 6 months	90	0	0	50	53.1	NS	NS	NS
CER1114581	benign	Benign	72	Male	NS	smoker, stopped over 5 months	3	0	0	84	30.5	NS	NS	NS
CER1114582	no baseline cancer	No baseline cancer	81	Male	NS	never smoker	0	0	0	11	2.1	NS	NS	NS
CER1114583	no baseline cancer	No baseline cancer	74	Male	NS	smoker, stopped over 6 months	10	0	1	13	9.4	NS	NS	NS
CER1114584	benign	Benign	47	Female	NS	never smoker	0	0	0	10	2.7	NS	NS	NS
CER1114585	benign	Benign	64	Male	NS	never smoker	0	0	0	10	4.8	NS	NS	NS
CER1114586	lung cancer	Adenocarcinoma	43	Male	IV	smoker	30	1	0	45	4.1	NS	NS	NS
CER1114587	no lung cancer, other cancer	No baseline cancer	76	Male	NS	smoker, stopped over 6 months	60	0	0	10	5.8	NS	breast cancer	1347
CER1114588	benign	Benign	73	Female	NS	smoker, stopped over 5 months	45	1	1	11	8.0	NS	NS	NS
CER1114589	lung cancer, other cancer	Squamous	84	Female	IIIb	smoker, stopped over 6 months	30	0	1	15	13.9	NS	bladder cancer	895
CER1114590	lung cancer	Adenocarcinoma	72	Female	IV	smoker, stopped over 6 months	50	0	1	14	2.8	NS	NS	NS
CER1114591	lung cancer	Adenocarcinoma	78	Female	IV	smoker, stopped over 5 months	10	0	1	10	3.8	NS	NS	NS
CER1114592	no baseline cancer	No baseline cancer	71	Female	NS	smoker, stopped over 6 months	10	0	1	10	4.5	NS	NS	NS
CER1114593	no lung cancer, other cancer	No baseline cancer	70	Male	NS	never smoker	0	0	0	10	1.4	metastatic melanoma, osteosarcoma, osteoarthritis	breast cancer, carcinoma	-9531, 1879, 2471
CER1114594	benign	Benign	51	Female	NS	smoker	40	0	0	10	3.2	NS	NS	NS
CER1114595	lung cancer	NSCL	80	Male	IV	smoker, stopped over 6 months	48	1	0	10	5.2	NS	NS	NS
CER1114596	no baseline cancer	No baseline cancer	61	Female	NS	smoker, stopped over 6 months	45	1	0	10	8.8	NS	NS	NS
CER1114597	no lung cancer, prior lung cancer, later lung cancer	No baseline cancer	57	Male	NS	smoker, stopped over 6 months	30	1	0	10	1.1	lung cancer, new primary NSCLC metastatic adenocarcinoma	NS	-811, 546
CER1114598	no baseline cancer	No baseline cancer	44	Male	NS	never smoker	2	1	0	15	2.1	NS	NS	NS
CER1114599	lung cancer	Adenocarcinoma	55	Female	IV	smoker	30	1	1	151	27.8	NS	NS	NS
CER1114600	met to lung, another primary cancer	Metastatic from another primary cancer	78	Female	IV	smoker, stopped over 6 months	1	0	0	20	4.1	breast cancer, osteosarcoma	NS	-1382
CER1114601	no baseline cancer	No baseline cancer	49	Female	NS	smoker	10	0	0	10	2.1	NS	NS	NS
CER1114602	lung cancer	Squamous	71	Male	IV	smoker, stopped over 6 months	120	0	0	10	10.4	NS	NS	NS
CER1114603	no baseline cancer	No baseline cancer	74	Male	NS	smoker	85	1	0	10	5.8	NS	NS	NS
CER1114604	met to lung, other cancer	Metastatic from another primary cancer	80	Female	IV	smoker, stopped over 5 months	15	0	0	40	8.4	breast cancer, osteosarcoma	NS	-1332

Patient ID	Patient type	Histological diagnosis	Age at diagnosis	Sex	Stage	Smoking status	Pack years	Autism spectrum disease (0=no, 1=yes)	CDP (0=no, 1=yes)	CRP (mg/L)	IL-6 (pg/mL)	Prior cancer type	Future cancer type	Time from blood draw to cancer diagnosis
CSP111423P	lung cancer	NSCLC-AdC	63	Male	IV	smoker, stopped over 5 months	27	0	0	25	81.3	N/A	N/A	N/A
CSP111434P	met to lung, another cancer	Metastatic from another primary cancer	33	Male	IV	never smoker	0	0	0	20	1	metastatic colorectal adenocarcinoma	N/A	~742
CSP111425P	met to lung, another cancer	Metastatic from another primary cancer	63	Female	IV	smoker	5	1	0	20	3.2	metastatic colorectal adenocarcinoma	N/A	~1,095, ~849
CSP111426P	no lung cancer, another cancer	no baseline cancer	61	Male	N/A	smoker, stopped over 5 months	33	1	1	20	2.9	N/A	prostate cancer operated	27
CSP111427P	lung cancer, another cancer	Adenocarcinoma	61	Female	IV	smoker	30	0	0	20	21.9	metastatic colorectal adenocarcinoma	N/A	~2,536
CSP111428P	lung cancer	Adenocarcinoma	66	Female	IV	smoker	60	0	1	21	3.2	N/A	N/A	N/A
CSP111429P	no baseline cancer	no baseline cancer	71	Male	N/A	smoker, stopped over 5 months	50	0	0	12	10.3	N/A	N/A	N/A
CSP111430P	no baseline cancer	no baseline cancer	50	Female	N/A	smoker	30	0	0	11	1.1	N/A	N/A	N/A
CSP111431P	met to lung	Metastatic from another primary cancer	30	Female	IV	smoker	21	0	0	20	13.9	N/A	N/A	N/A
CSP111432P	benign	benign	63	Female	N/A	smoker, stopped over 5 months	12	0	0	20	5.5	N/A	N/A	N/A
CSP111433P	no lung cancer, another cancer	no baseline cancer	75	Male	N/A	smoker, stopped over 5 months	30	0	0	20	2.9	breast cancer operated	N/A	~2,855
CSP111434P	no lung cancer, other lung cancer, later lung cancer	no baseline cancer	62	Female	N/A	smoker	40	0	0	20	3.9	breast cancer operated	NSCLC, adenocarcinoma, stage I	~2,547, ~130
CSP111435P	no baseline cancer	no baseline cancer	57	Male	N/A	never smoker	0	1	0	20	2.9	N/A	N/A	N/A
CSP111436P	no baseline cancer	no baseline cancer	70	Male	N/A	smoker, stopped over 5 months	102	1	1	12.8	21.3	N/A	N/A	N/A
CSP111437P	no baseline cancer	no baseline cancer	77	Female	N/A	smoker, stopped over 5 months	20	0	1	20	3.2	N/A	N/A	N/A
CSP111438P	no lung cancer, other cancer	benign	50	Female	N/A	smoker, stopped over 5 months	30	0	0	20	1.1	N/A	breast cancer operated	641
CSP111439P	no baseline cancer	no baseline cancer	79	Male	N/A	smoker, stopped over 5 months	12	0	0	20	3.4	N/A	N/A	N/A
CSP111500P	lung cancer, another cancer	lymphoma	75	Female	IV	smoker	25	0	0	71	78.7	metastatic colorectal adenocarcinoma	N/A	~4,312
CSP111501P	lung cancer, other cancer	Adenocarcinoma	76	Female	IV	smoker	27	0	1	33	110	N/A	breast cancer	~2,521
CSP111502P	no baseline cancer	no baseline cancer	55	Female	N/A	smoker	60	0	1	20	3.3	N/A	N/A	N/A
CSP111503P	no lung cancer, other cancer	benign	65	Male	N/A	smoker, stopped over 5 months	15	1	0	32	12.3	metastatic colorectal adenocarcinoma	metastatic colorectal adenocarcinoma	~2,575, ~602
CSP111504P	lung cancer	Adenocarcinoma	58	Female	N/A	smoker	14	0	0	22	6.4	N/A	N/A	N/A
CSP111505P	lung cancer	Adenocarcinoma	57	Female	IV	smoker	43	1	0	35	19.3	N/A	N/A	N/A
CSP111506P	lung cancer, other cancer	Squamous	76	Male	N/A	smoker, stopped over 5 months	70	0	0	10	14	N/A	N/A	~1,665

Patient ID	Patient type	Histological diagnosis	Age at diagnosis	Sex	Stage	Smoking status	Past years	Autoimmune disease (Yes, 1=Yes)	CSPD (No=0, 1=Yes)	CSP mg/L	IL-6 (pg/ml)	Prior cancer type	Future cancer type	Time from blood draw to cancer diagnosis
CSP111051P	no lung cancer, other cancer	no baseline cancer	70	female	NA	never smoker	0	1	0	10	7	breast cell carcinoma	breast cell carcinoma	-1458, 1319
CSP111050P	benign	benign	24	Male	NA	smoker	5	0	0	10	3.8	NA	NA	NA
CSP111050P	no baseline cancer	no baseline cancer	70	female	NA	smoker, stopped over 5 months	13	1	0	10	3.2	NA	NA	NA
CSP111051P	lung cancer, other cancer	adenocarcinoma	85	female	NA	smoker, stopped over 5 months	10	0	0	10	5.2	breast cell carcinoma	NA	-1521, -1195
CSP111051P	no lung cancer, other cancer	no baseline cancer	81	female	NA	smoker, stopped over 5 months	20	1	0	10	5.4	adenocarcinoma, operated	NA	-1128
CSP111051P	no baseline cancer	no baseline cancer	89	Male	NA	never smoker	0	0	0	10	2.4	NA	NA	NA
CSP111051P	no lung cancer, other cancer	no baseline cancer	51	female	NA	smoker, stopped over 5 months	40	1	0	10	1.6	epidermal cancer, operated	epidermal cancer	-2545, 1779
CSP111051P	no lung cancer, other cancer	no baseline cancer	74	Male	NA	smoker	70	0	0	11	6.2	renal cancer, operated	renal cancer, operated	-5915, 2179
CSP111051P	benign	benign	56	Male	NA	smoker	23	1	0	10	1.9	NA	NA	NA
CSP111051P	lung cancer, other cancer	adenocarcinoma	58	female	IV	smoker	13	0	1	84	21.5	adenocarcinoma, breast cell carcinoma	NA	-13150, -325
CSP111051P	no lung cancer, other cancer	benign	75	female	NA	smoker, stopped over 5 months	11	0	0	10	4.8	thyroid cancer, operated, rectal cancer, operated	thyroid cancer, operated, rectal cancer, operated	-2580, -410, 83
CSP111051P	lung cancer	adenocarcinoma	88	female	NA	smoker	52	1	0	84	16.7	NA	NA	NA
CSP111051P	benign	benign	66	Male	NA	smoker, stopped over 5 months	30	1	0	40	5.5	NA	NA	NA
CSP111052P	benign	benign	43	Male	NA	never smoker	0	0	0	10	0.95	NA	NA	NA
CSP111051P	benign	benign	41	female	NA	smoker, stopped over 5 months	20	1	0	10	1.1	NA	NA	NA
CSP111052P	no lung cancer, other cancer	no baseline cancer	87	female	NA	smoker, stopped over 5 months	20	1	1	10	1.3	breast cell carcinoma	chronic lymphocytic leukaemia	-3507, 1403
CSP111051P	benign	benign	17	Male	NA	smoker, stopped over 5 months	12	0	0	10	1.8	NA	NA	NA
CSP111051P	no lung cancer, later lung cancer	no baseline cancer	67	Male	NA	smoker	15	1	0	11	5.7	NA	rectal cancer, stages I, colonoscopy	253
CSP111051P	benign	benign	37	Male	NA	never smoker	0	1	0	10	2.9	NA	NA	NA
CSP111052P	lung cancer	metastatic	76	Male	IV	smoker	70	0	0	58	23.3	NA	NA	NA
CSP111051P	no lung cancer, other cancer	benign	75	female	NA	smoker, stopped over 5 months	3	0	0	55	16.1	NA	breast cell carcinoma	34
CSP111051P	no baseline cancer	no baseline cancer	66	female	NA	never smoker	0	0	0	10	3.9	NA	NA	NA
CSP111051P	metastatic lung cancer, primary cancer	metastasis from smoker primary cancer	57	Male	IV	smoker, stopped over 5 months	60	0	1	285	45.6	NA	NA	NA
CSP111053P	no lung cancer, later lung cancer	benign	74	Male	NA	smoker	60	0	1	10	16.7	lung cancer, metastatic, prostate	lung cancer	1993
CSP111053P	lung cancer	SCLC	77	Male	IV	smoker, stopped over 5 months	50	0	0	12	16.7	NA	NA	NA
CSP111053P	no baseline cancer	no baseline cancer	50	female	NA	smoker, stopped over 5 months	40	0	0	10	1.1	NA	NA	NA

Patient ID	Patient type	Histological diagnosis	Age at diagnosis	Sex	Stage	Smoking status	Pack years	Autism spectrum disease (No./Yes)	CPEP (No./Yes)	CPEP (ng/L)	IL-6 (pg/mL)	Prior cancer type	Future cancer type	Time from blood draw to cancer diagnosis
C08L00531F	lung cancer, later lung cancer	adenocarcinoma	75	Female	IA	smoker	40	0	0	10	3.2	N/A	lung cancer, NSCLC, neuroendocrine	2132
C08L00534F	Benign	Benign	49	Female	N/A	smoker, stopped smoker 6 months	25	0	0	11	2.5	N/A	N/A	N/A
C08L00535F	no lung cancer, other cancer	Benign	63	Male	N/A	smoker, stopped over 6 months	43	1	0	10	3.2	N/A	bladder carcinoma	1124
C08L00536F	no lung cancer, other cancer	No baseline cancer	61	Female	N/A	never smoker	0	0	0	10	5.2	colon cancer, operated adenocarcinoma	N/A	214
C08L00537F	no baseline cancer	No baseline cancer	65	Female	N/A	smoker, stopped over 6 months	15	1	1	10	3.7	N/A	N/A	N/A
C08L00538F	lung cancer, other cancer	adenocarcinoma	78	Female	IV	smoker, stopped over 6 months	65	0	1	10	5	adenocarcinoma in situ, operated	N/A	13886
C08L00539F	Benign	Benign	77	Male	N/A	smoker, stopped over 6 months	20	0	0	10	3.3	N/A	N/A	N/A
C08L00540F	Benign	Benign	70	Male	N/A	smoker, stopped over 6 months	50	0	0	10	4.9	N/A	N/A	N/A
C08L00541F	no baseline cancer	No baseline cancer	53	Male	N/A	smoker	25	0	0	10	3	N/A	N/A	N/A
C08L00542F	metastatic from another primary cancer	Metastatic from another primary cancer	65	Male	IV	smoker	75	0	0	10	13.3	obovary cancer, operated	N/A	1452
C08L00543F	lung cancer	adenocarcinoma	64	Female	IV	smoker	30	0	0	17	22.5	N/A	N/A	N/A
C08L00544F	lung cancer	adenocarcinoma	60	Female	IIA	smoker	60	0	1	10	3.1	N/A	N/A	N/A
C08L00545F	no lung cancer, other cancer	No baseline cancer	57	Female	N/A	smoker, stopped over 6 months	25	1	0	10	3.4	breast cancer, operated	N/A	1909
C08L00546F	no lung cancer, other cancer	No baseline cancer	66	Female	N/A	smoker, stopped over 6 months	15	1	1	42	394	non-invasive squamous basaloid papillary lymphoma of the penis	N/A	2813
C08L00547F	no baseline cancer	No baseline cancer	56	Female	N/A	smoker	75	0	1	10	2.5	N/A	N/A	N/A
C08L00548F	no baseline cancer	No baseline cancer	72	Male	N/A	smoker, stopped over 6 months	20	0	0	10	5.7	N/A	N/A	N/A
C08L00549F	no lung cancer, other cancer	No baseline cancer	65	Male	N/A	never smoker	0	0	0	18	13	N/A	T cell lymphoma	2457
C08L00550F	lung cancer, other cancer	adenocarcinoma	72	Male	IV	smoker	100	0	0	10	1.3	squamous cell carcinoma, operated	N/A	1352
C08L00551F	no baseline cancer	No baseline cancer	60	Male	N/A	smoker, stopped over 6 months	23	0	1	18	1.5	N/A	N/A	N/A
C08L00552F	no lung cancer, later lung cancer	Benign	56	Male	N/A	smoker	65	0	0	12	135.2	lung cancer, operated	NSC, stage II, squamous, squamous cancer	1987 & 1346
C08L00553F	no lung cancer, other cancer	Benign	69	Male	N/A	smoker	55	1	0	19	3	basal cell carcinoma	N/A	1796
C08L00554F	no lung cancer, other cancer	Benign	56	Male	N/A	never smoker	0	1	0	10	1.8	testicular cancer, operated	N/A	1207
C08L00555F	no lung cancer, other cancer	No baseline cancer	50	Female	N/A	smoker, stopped over 6 months	10	0	1	10	9.2	N/A	uterine cancer	320

Patient ID	Patient type	Histological diagnosis	Age at diagnosis	Sex	Stage	Smoking status	Pack years	Autoimmune disease (0=No, 1=Yes)	CORD (0=No, 1=Yes)	CSP (mg/L)	IL-2 (pg/mL)	Prior cancer type	Future cancer type	Time from blood draw to cancer diagnosis
COP10256P	No baseline cancer	No baseline cancer	82	Female	N/A	never smoker	0	0	0	20	2.3	N/A	N/A	N/A
COP10257P	No lung cancer, other cancer	No baseline cancer	79	Male	N/A	smoker, stopped over 6 months	19	1	0	30	1	malignant melanoma, squamous	N/A	-906, -25
COP10258P	lung cancer	Squamous	67	Male	IIA	smoker	60	1	1	20	N/A	N/A	N/A	N/A
COP10259P	lung cancer, prior lung cancer	Adenocarcinoma	71	Female	IV	smoker	150	1	0	20	8	lung cancer, NSCLL, operated adenocarcinoma, recurrent in 2020	N/A	-1061, -1079
COP10260P	No baseline cancer	No baseline cancer	15	Male	N/A	never smoker	0	0	0	21	1.9	N/A	N/A	N/A
COP10261P	benign	Benign	48	Male	N/A	smoker	35	1	0	20	24.1	N/A	N/A	N/A
COP10262P	No lung cancer, other cancer	Benign	60	Male	N/A	smoker, stopped over 6 months	50	1	0	20	2.5	prostate cancer, adenocarcinoma	N/A	-1213
COP10263P	No lung cancer, other cancer	No baseline cancer	30	Female	N/A	never smoker	0	0	0	20	0.5	lymphoma	N/A	-7287
COP10264P	No lung cancer, other cancer	No baseline cancer	65	Female	N/A	smoker, stopped over 6 months	15	0	1	20	23.1	N/A	squamous cell carcinoma	1485
COP10265P	lung cancer	Adenocarcinoma	55	Female	IIIA	smoker, stopped over 6 months	40	0	0	30	14.9	N/A	N/A	N/A
COP10266P	lung cancer	Squamous	71	Male	IV	smoker, stopped over 6 months	50	0	0	29	22.5	N/A	N/A	N/A
COP10267P	No baseline cancer	No baseline cancer	60	Male	N/A	smoker, stopped over 6 months	60	1	0	20	13.2	N/A	N/A	N/A
COP10268P	benign	Benign	47	Female	N/A	smoker	30	0	0	29	35.4	N/A	N/A	N/A
COP10269P	lung cancer	Squamous	74	Female	IA	smoker, stopped over 6 months	50	1	0	20	1.5	N/A	N/A	N/A
COP10270P	No lung cancer, prior lung cancer	No baseline cancer	56	Female	N/A	smoker	43	0	0	20	7.1	N/A	lung cancer, SCC	1016
COP10271P	lung cancer	Adenocarcinoma	67	Male	IA	smoker, stopped over 6 months	75	0	1	34	10	N/A	N/A	N/A
COP10272P	No baseline cancer	No baseline cancer	57	Female	N/A	smoker, stopped over 6 months	0	1	0	27	4.8	N/A	N/A	N/A
COP10273P	No baseline cancer	No baseline cancer	50	Male	N/A	smoker	40	0	0	20	2.7	N/A	N/A	N/A
COP10274P	benign	Benign	49	Female	N/A	smoker	25	0	1	20	1.9	N/A	N/A	N/A
COP10275P	No lung cancer, other cancer	No baseline cancer	50	Female	N/A	smoker	45	0	0	20	0.7	pancreatic disease of the digestive	N/A	-49
COP10276P	No baseline cancer	No baseline cancer	51	Female	N/A	smoker, stopped over 6 months	25	0	0	20	2	N/A	N/A	N/A
COP10277P	No baseline cancer	No baseline cancer	67	Male	N/A	never smoker	0	0	0	50	44.1	N/A	N/A	N/A
COP10278P	lung cancer	Adenocarcinoma	70	Female	IIIA	smoker, stopped over 6 months	50	0	0	13	4.3	N/A	N/A	N/A
COP10280P	No baseline cancer	No baseline cancer	63	Male	N/A	never smoker	0	0	0	20	0.87	N/A	N/A	N/A
COP10282P	benign	Benign	58	Female	N/A	smoker	46	0	0	20	2.18	N/A	N/A	N/A
COP10283P	benign	Benign	41	Male	N/A	smoker, stopped over 6 months	10	0	0	20	1	N/A	N/A	N/A
COP10284P	No baseline cancer	No baseline cancer	30	Male	N/A	smoker, stopped over 6 months	10	1	0	20	0.78	N/A	N/A	N/A

Patient ID	Patient type	Histological diagnosis	Age at diagnosis	Sex	Stage	Smoking status	Pack years	Autoimmune disease (date, 1=Yes)	COPD (date, 1=Yes)	CRP (mg/L)	PLC (pg/mL)	Prior cancer type	Future cancer type	Time from blood draw to cancer diagnosis
C08LL0304P	lung cancer	adenocarcinoma	64	Male	IB	smoker, stopped over 6 months	25	0	1	10	3.6	N/A	N/A	N/A
C08LL0305P	lung cancer, other cancer	adenocarcinoma	76	Female	IBB	smoker	45	0	0	10	5.9	basal cell carcinoma	N/A	-751
C08LL0306P	no baseline cancer	no baseline cancer	73	Female	IIA	smoker, stopped over 6 months	20	1	0	21	4	N/A	N/A	N/A
C08LL0307P	lung cancer	adenocarcinoma	79	Male	IV	smoker, stopped over 6 months	55	0	0	21	9.7	N/A	N/A	N/A
C08LL0308P	lung cancer, other cancer	SCLC	61	Female	IV	smoker, stopped over 6 months	35	0	0	10	7.7	uterine cancer, operated	N/A	-10978
C08LL0309P	benign	benign	58	Male	IIA	smoker, stopped over 6 months	30	0	0	10	1.9	N/A	N/A	N/A
C08LL0310P	benign	benign	42	Female	IIA	never smoker	0	1	0	20	1.6	N/A	N/A	N/A
C08LL0311P	metastatic, another primary cancer	metastatic from another primary cancer	85	Female	IV	never smoker	0	0	0	12	20.6	breast cancer, operated	N/A	-7794, -9482, -4882
C08LL0312P	benign	benign	67	Male	IIA	smoker, stopped over 6 months	40	1	0	10	1.0	N/A	N/A	N/A
C08LL0313P	no baseline cancer	no baseline cancer	66	Male	IIA	never smoker	0	0	0	10	0.21	N/A	N/A	N/A
C08LL0314P	lung cancer, SCLC	SCLC	86	Female	IBB	smoker	30	0	1	10	6.96	N/A	N/A	N/A
C08LL0315P	no baseline cancer	no baseline cancer	57	Female	IIA	smoker, stopped over 6 months	48	1	0	10	2.7	N/A	N/A	N/A
C08LL0316P	no baseline cancer	no baseline cancer	51	Male	IIA	never smoker	0	1	0	10	3.9	N/A	N/A	N/A
C08LL0317P	benign	benign	25	Female	IIA	smoker	5	2	0	10	3.3	N/A	N/A	N/A
C08LL0318P	no lung cancer, other cancer	no baseline cancer	73	Female	IIA	smoker, stopped over 6 months	40	1	1	10	3.9	breast cancer, operated	N/A	-1254
C08LL0319P	no lung cancer, other cancer	no baseline cancer	64	Female	IIA	smoker, stopped over 6 months	10	0	0	10	0.4	breast cancer, operated	basal cell carcinoma	-1307, -2495
C08LL0320P	no baseline cancer	no baseline cancer	75	Male	IIA	smoker, stopped over 6 months	N/A	0	0	10	1.9	N/A	N/A	N/A
C08LL0321P	no baseline cancer	no baseline cancer	56	Male	IIA	smoker, stopped over 6 months	30	0	0	10	1.6	N/A	N/A	N/A
C08LL0322P	no lung cancer, other cancer	no baseline cancer	80	Female	IIA	smoker, stopped over 6 months	N/A	1	0	10	5.7	basal cell carcinoma	N/A	-1791, -1112
C08LL0323P	lung cancer	benign	67	Male	IIA	never smoker	0	0	0	10	4.1	N/A	N/A	N/A
C08LL0324P	lung cancer, prior lung cancer	adenocarcinoma	59	Male	IBB	smoker	30	0	1	88	794	N/A	N/A	N/A
C08LL0325P	lung cancer, prior lung cancer	adenocarcinoma	64	Male	IA	smoker	150	0	0	10	1.1	lung cancer, NSCLC operated before, recurrent in 2012	N/A	-1889, -1104
C08LL0326P	no baseline cancer	no baseline cancer	60	Male	IIA	smoker	55	1	0	41	47.3	N/A	N/A	N/A
C08LL0327P	benign	benign	60	Male	IIA	smoker, stopped over 6 months	30	0	1	15	766	N/A	N/A	N/A
C08LL0328P	no baseline cancer	no baseline cancer	68	Male	IIA	smoker, stopped over 6 months	40	0	0	10	1.7	N/A	N/A	N/A
C08LL0329P	lung cancer, other cancer	adenocarcinoma	76	Male	IV	smoker	30	0	0	12	15.3	prostate cancer, adeno	N/A	-1453

Patient ID	Patient type	Histopathological diagnosis	Age at diagnosis	Sex	Stage	Smoking status	Pack years	Autoimmune disease (Yes/No)	CEP (mg/L)	HLG (mg/dL)	Prior cancer type	Future cancer type	Time from blood draw to cancer diagnosis
C0P1U0612P	metastatic, other cancer	Metastasis from another primary cancer	56	Male	N	Smoker	40	0	0	5.4	Squamous squamous metastatic, operated	N/A	525
C0P1U0613P	no lung cancer, other cancer	No baseline cancer	58	Male	N/A	Never smoker	0	1	0	5.5	N/A	Basal cell carcinoma, operated	53, 2343
C0P1U0614P	no lung cancer, other cancer	No baseline cancer	58	Female	N/A	Smoker, stopped over 6 months	20	0	0	2.6	Basal cell carcinoma, operated	N/A	2214
C0P1U0615P	benign	Benign	70	Female	N/A	Never smoker	0	0	1	2.7	N/A	N/A	N/A
C0P1U0616P	lung cancer	Adenocarcinoma	55	Male	I	Smoker, stopped over 6 months	25	0	1	1.1	N/A	N/A	N/A
C0P1U0617P	no lung cancer, other cancer	No baseline cancer	75	Male	N/A	Smoker, stopped over 6 months	N/A	0	0	10	13.4	Bladder cancer	2355
C0P1U0618P	no lung cancer, other cancer	No baseline cancer	26	Female	N/A	Smoker, stopped over 6 months	15	1	0	1.4	Basal cell carcinoma	Basal cell carcinoma x 2	424, 480, 2453
C0P1U0619P	no lung cancer, other lung cancer	Benign	51	Female	N/A	Smoker, stopped over 6 months	15	0	0	N/A	N/A	N/A	236
C0P1U0620P	no lung cancer, other cancer	No baseline cancer	54	Female	N/A	Smoker	30	1	0	1.0	Basal cell carcinoma, operated	N/A	4802
C0P1U0621P	no baseline cancer	No baseline cancer	38	Female	N/A	Smoker	20	0	0	19.4	N/A	N/A	N/A
C0P1U0622P	lung cancer, other cancer	Squamous	63	Male	III	Smoker, stopped over 6 months	50	0	0	13	N/A	Basal cell carcinoma, squamous cell carcinoma	40, 270
C0P1U0623P	benign	Benign	63	Male	N/A	Smoker, stopped over 6 months	41	0	0	1.0	N/A	N/A	N/A
C0P1U0624P	no baseline cancer	No baseline cancer	40	Male	N/A	Smoker, stopped over 6 months	1	0	0	2.1	N/A	N/A	N/A
C0P1U0625P	lung cancer	Metastatic adenocarcinoma	45	Male	IV	Smoker	50	0	1	9.3	N/A	N/A	N/A
C0P1U0626P	lung cancer	Squamous	70	Male	IA	Smoker	100	0	0	2.1	N/A	N/A	N/A
C0P1U0627P	lung cancer	Metastatic	71	Male	III	Smoker	60	0	1	600	N/A	N/A	N/A
C0P1U0628P	lung cancer	Squamous	35	Female	I	Smoker, stopped over 6 months	40	0	0	12.3	N/A	N/A	N/A
C0P1U0629P	lung cancer, other cancer	Adenocarcinoma	90	Male	I	Smoker, stopped over 6 months	20	0	0	145	COLON CANCER, OPERATED 2010, SQUAMOUS CELL CARCINOMA, BLADDER CELL CARCINOMA	N/A	4511, 2355, 534
C0P1U0630P	benign	Benign	39	Male	N/A	Never smoker	0	0	0	1.0	N/A	N/A	N/A
C0P1U0631P	benign	Benign	49	Male	N/A	Smoker	100	0	1	2.5	N/A	N/A	N/A
C0P1U0632P	no lung cancer, other cancer	No baseline cancer	35	Female	N/A	Smoker, stopped over 6 months	5	0	0	1.0	COLON CANCER, OPERATED 2010	N/A	22
C0P1U0633P	lung cancer	Adenocarcinoma	60	Male	N/A	Smoker	40	1	0	7.0	N/A	N/A	N/A
C0P1U0634P	benign	Benign	37	Male	N/A	Smoker	25	1	0	7.1	N/A	N/A	N/A
C0P1U0635P	benign	Benign	61	Female	N/A	Smoker, stopped over 6 months	25	0	0	46.9	N/A	N/A	N/A
C0P1U0636P	lung cancer	Adenocarcinoma	74	Female	III	Smoker, stopped over 6 months	25	0	0	11.1	N/A	N/A	N/A
C0P1U0637P	lung cancer	Squamous	69	Male	IV	Smoker	50	1	0	14.8	N/A	N/A	N/A

Patient ID	Patient type	Histological diagnosis	Age at diagnosis	Sex	Stage	Smoking status	Past years	Autoimmune disease (I=No, I=Yes)	CCP (I=No, I=Yes)	CRP (mg/L)	IL-6 (pg/ml)	Prior cancer type	Prior cancer type	Time from blood draw to cancer diagnosis
0021114520	Long-term	Adenocarcinoma	65	Male	II	Smoker, stopped under 6 months	20	0	0	127	102.8	NA	NA	NA
0021114520	Long-term	Adenocarcinoma	71	Male	II	Smoker, stopped over 6 months	40	0	1	68	11.6	NA	NA	NA

Supplementary Table 3. Summary of LUCAS cohort and genomic analyses* (CONTINUED)

Survival status	Days since alive	cDNA extraction batch	Sequencing library batch	Sequencing machine	Flow cell	Sequencing approach	Read length	Number of sequenced fragments	Number of sequenced reads	Percent of D20 lanes	Percent of Q20 bases	Reads mapped to genome	Whole genome coverage	Estimated 500x-1G level	Inter-tumor fraction	Single molecule C.A. frequency changes	Single molecule C.A. frequency (per Mb of exome)	Regional difference in single molecule C.A. frequency (per Mb of exome)	QEM098 score	QEM098 score	QEM098 score
Alive	2779	1	14	1	W65	PureSelect	100	36,437,817	76,673,634	36.09%	93.92%	70,587,461	2.3	1.24	0.61%	1.688	14.39	-0.13	0.59	-	-
Alive	2779	1	2	2	W65	PureSelect	100	40,130,400	80,240,800	36.93%	93.73%	79,336,000	2.4	1.36	0.63%	1.732	13.80	-1.37	0.59	0.01	0.20
Alive	2779	1	72	3	W65	PureSelect	100	34,709,261	69,566,882	34.77%	91.73%	69,245,513	2.0	1.15	0.60%	1.706	12.95	-4.34	0.59	-	-
Deceased	2807	1	2	2	W65	PureSelect	100	59,732,841	77,522,282	36.70%	93.76%	71,582,039	2.3	1.36	0.63%	1.695	12.95	-2.73	0.54	-	-
Deceased	2903	1	3	3	W65	PureSelect	100	42,636,593	85,373,106	36.70%	93.80%	85,607,595	2.7	1.19	0.71%	1.760	15.04	-4.25	0.80	0.06	0.15
Alive	2779	4	4	2	W65	PureSelect	100	40,649,504	81,299,016	37.79%	92.15%	71,633,951	2.4	1.03	0.73%	1.555	34.97	1.66	0.61	-	-
Alive	2779	1	5	3	W65	PureSelect	100	40,146,951	81,293,902	37.41%	92.49%	71,670,951	2.4	1.31	0.80%	2.238	14.52	-1.34	0.59	0.00	0.15
Deceased	487	1	5	3	W65	PureSelect	100	59,832,153	79,706,170	36.45%	93.40%	79,222,217	2.4	1.30	0.80%	1.857	14.59	-1.39	0.49	-	-
Deceased	187	1	4	2	W65	PureSelect	100	40,824,893	80,246,480	33.15%	90.17%	77,521,531	2.4	1.04	0.60%	1.888	34.69	-4.71	0.88	-	-
Deceased	583	1	4	2	W65	PureSelect	100	54,489,901	80,948,482	34.77%	92.67%	79,525,513	2.4	1.16	0.62%	1.817	15.02	-0.24	0.51	-	-
Alive	2779	1	6	2	W65	PureSelect	100	57,943,983	67,831,610	37.01%	94.49%	69,327,532	2.0	1.22	0.64%	1.096	13.97	-1.39	0.22	-	-
Deceased	425	1	2	2	W65	PureSelect	100	42,477,532	84,955,104	39.33%	93.13%	79,371,946	2.6	1.18	0.62%	1.871	14.15	0.02	0.74	0.82	0.24
Deceased	97	2	3	3	W65	PureSelect	100	36,806,392	73,612,784	36.77%	93.15%	69,634,351	2.2	1.23	0.60%	1.607	14.22	10.35	1.60	1.00	1.40
Alive	2779	2	5	2	W65	PureSelect	100	36,703,261	73,406,524	36.79%	93.09%	69,668,661	2.0	1.24	0.63%	1.537	13.78	0.34	0.60	-	-
Deceased	139	2	34	2	W65	PureSelect	100	52,277,842	64,395,204	37.80%	95.29%	59,827,682	2.6	1.28	0.23%	2.492	15.48	0.47	0.79	-	-
Deceased	139	2	13	1	W65	PureSelect	100	59,358,513	73,111,850	34.91%	91.49%	69,444,002	2.2	1.12	0.62%	1.887	41.96	-1.08	0.47	-	-
Alive	2774	2	18	3	W65	PureSelect	100	41,993,330	83,987,760	35.96%	91.74%	79,538,017	2.4	1.16	0.60%	1.180	11.99	-4.00	0.32	0.01	0.18
Alive	2774	2	45	3	W65	PureSelect	100	39,750,934	71,520,068	34.73%	91.35%	61,147,975	2.0	1.24	0.43%	1.382	23.82	-5.57	0.65	-	-
Alive	2774	2	4	2	W65	PureSelect	100	34,070,761	68,141,402	35.30%	92.70%	67,460,116	2.0	1.21	1.13%	1.435	13.17	-1.11	0.46	0.26	0.33
Alive	2774	2	28	2	W65	PureSelect	100	59,112,950	78,942,438	34.24%	92.13%	69,133,666	2.2	1.38	0.89%	3.753	27.63	-0.27	0.59	-	-
Deceased	2735	2	6	2	W65	PureSelect	100	55,843,973	77,021,946	37.00%	91.70%	69,549,698	2.1	1.17	0.70%	1.557	15.74	-0.44	0.59	-	-
Deceased	58	2	14	1	W65	PureSelect	100	55,490,267	70,600,414	34.51%	91.77%	61,700,594	2.0	1.24	0.60%	1.046	16.01	1.69	0.80	-	-
Deceased	58	2	6	3	W65	PureSelect	100	42,647,571	85,295,142	39.20%	93.27%	77,638,740	2.5	1.37	0.36%	2.665	17.11	1.39	0.78	1.00	0.80
Alive	2772	2	42	3	W65	PureSelect	100	37,708,241	75,416,342	36.30%	94.25%	69,444,706	2.3	1.37	0.34%	1.805	14.72	-1.03	0.32	-	-
Alive	2772	3	34	2	W65	PureSelect	100	36,689,514	73,379,028	36.79%	91.05%	66,512,118	2.1	1.27	0.73%	3.136	21.88	-2.31	0.58	-	-
Alive	2771	3	34	4	W65	PureSelect	100	53,154,468	69,540,996	34.36%	80.75%	59,536,410	2.4	1.52	0.67%	1.679	15.05	-3.36	0.33	0.50	0.24
Alive	2771	3	17	1	W65	PureSelect	100	46,361,864	92,723,728	34.95%	91.13%	81,141,331	2.7	1.17	0.51%	1.708	18.03	-4.33	0.89	-	-
Deceased	1899	3	48	3	W65	PureSelect	100	42,898,856	85,797,712	35.13%	91.19%	77,632,326	2.5	1.62	1.16%	2.794	17.82	-4.64	0.74	-	-
Deceased	1899	3	44	3	W65	PureSelect	100	37,604,989	75,198,996	35.89%	91.25%	67,590,391	2.2	1.79	0.81%	2.106	15.09	-1.49	0.65	-	-
Alive	2771	3	6	2	W65	PureSelect	100	41,172,820	82,345,760	35.20%	91.45%	76,405,325	2.5	1.26	1.03%	2.260	14.45	1.18	0.78	-	-
Deceased	2773	3	12	2	W65	PureSelect	100	59,042,462	73,095,924	35.41%	92.54%	71,564,290	2.3	1.57	0.78%	2.604	15.44	-2.39	0.79	1.00	0.84
Deceased	279	3	12	3	W65	PureSelect	100	40,132,155	80,264,312	35.70%	92.79%	79,827,517	2.4	1.59	0.82%	1.835	15.38	-0.49	0.59	0.00	0.33
Alive	2779	3	1	4	W65	PureSelect	100	59,646,280	92,186,560	36.41%	94.30%	61,845,516	2.0	1.29	0.44%	3.484	28.15	-4.62	0.80	0.61	0.45
Alive	2779	3	2	2	W65	PureSelect	100	36,632,869	73,265,738	35.84%	92.51%	71,631,092	2.3	1.74	0.32%	2.118	14.40	2.36	0.80	0.06	0.45
Deceased	52	3	14	4	W65	PureSelect	100	41,435,671	82,870,442	38.81%	93.99%	77,009,218	2.5	1.43	4.63%	1.967	13.02	0.46	0.74	0.14	0.44
Deceased	59	3	15	3	W65	PureSelect	100	42,038,579	84,077,158	35.15%	92.45%	77,644,041	2.3	1.78	0.33%	2.182	13.99	0.26	0.67	0.94	0.22
Alive	2770	4	5	2	W65	PureSelect	100	58,939,142	77,871,036	34.77%	93.89%	72,141,036	2.4	1.57	0.80%	1.895	15.02	0.29	0.67	-	-
Deceased	42	4	1	6	W65	PureSelect	100	51,999,711	63,996,482	35.47%	92.67%	59,115,519	1.8	1.17	0.63%	2.267	24.38	3.43	0.93	0.89	0.96
Alive	2787	5	4	2	W65	PureSelect	100	41,149,289	82,298,578	34.99%	91.15%	77,704,519	2.5	1.74	0.63%	1.940	15.33	-1.90	0.39	0.22	0.31
Deceased	395	5	13	3	W65	PureSelect	100	59,822,723	73,645,446	34.40%	90.27%	69,310,764	2.3	1.59	0.32%	1.997	14.16	1.32	0.84	-	-
Deceased	836	5	25	3	W65	PureSelect	100	32,008,058	64,016,116	30.56%	80.24%	60,232,215	2.3	1.99	0.39%	1.660	22.51	6.32	0.70	-	-
Deceased	28	5	26	1	W65	PureSelect	100	37,262,553	74,525,106	35.79%	93.96%	68,633,544	2.2	1.49	3.32%	2.694	15.42	-3.23	0.18	-	-
Alive	2768	5	9	3	W65	PureSelect	100	52,433,718	64,537,432	35.99%	92.14%	60,155,016	1.9	1.60	0.74%	1.596	14.01	-0.18	0.60	0.12	0.39
Alive	2765	5	6	4	W65	PureSelect	100	37,554,263	65,108,730	35.19%	92.89%	59,324,417	1.7	1.36	1.40%	1.396	14.43	-1.02	0.39	0.00	0.17
Deceased	734	5	7	2	W65	PureSelect	100	43,278,331	86,556,662	33.61%	90.17%	76,446,516	2.4	1.72	0.64%	3.596	27.65	0.36	0.71	0.36	0.53
Alive	2763	5	8	3	W65	PureSelect	100	41,604,893	83,212,776	33.90%	90.49%	69,687,768	2.2	1.38	0.71%	3.600	21.99	-0.08	0.63	-	-
Deceased	821	5	8	2	W65	PureSelect	100	41,102,897	82,205,714	35.13%	92.75%	71,738,172	2.3	1.40	0.52%	1.640	13.02	-0.53	0.55	-	-
Deceased	2718	5	21	3	W65	PureSelect	100	53,867,273	107,734,546	34.20%	90.95%	90,939,431	3.2	1.43	0.85%	4.140	22.94	-5.09	0.62	-	-
Deceased	79	5	20	2	W65	PureSelect	100	53,539,415	65,136,830	35.90%	92.69%	59,036,438	1.8	1.46	1.46%	2.630	24.42	0.68	0.71	0.84	0.64
Deceased	660	5	24	2	W65	PureSelect	100	58,821,320	73,761,640	37.64%	95.12%	71,411,688	1.6	1.45	1.02%	3.021	14.13	-7.12	0.62	0.61	0.31

Survival status	Days since	cDNA extraction batch	Genomic library batch	Sequencing machine	Flow cell	Lane of flow cell	Sequencing approach	Read length	Number of sequenced fragments	Number of sequenced reads	Percent of Q30 reads	Percent of Q30 bases	Reads mapped to genome	Whole genome coverage	Estimated 800-100 bp fraction	Identified variant fraction	Single molecule CNA changes	Single molecule CNA frequency (per Mb of exome)	Aggregat difference in single molecule CNA frequency (per Mb of exome)	OSMRR of 250K score	OSMRR score	
Alive	2760	6	11	3	H	6	WGS	100	17,906,539	64,813,000	93.27%	56,536,331	56,536,331	1.8	1.56	1.47%	1.54%	14.30	-1.81	0.31	-	-
Alive	2760	6	4	2	A	4	WGS	100	17,954,517	65,158,034	93.00%	56,942,135	56,942,135	2.0	1.36	0.77%	2.07%	17.75	6.34	0.94	0.07	0.43
Alive	2760	6	13	3	F	3	WGS	100	19,008,135	70,600,670	96.73%	64,427,487	64,427,487	2.1	1.47	0.43%	1.60%	14.30	0.55	0.21	-	-
Deceased	2760	6	4	2	A	4	WGS	100	16,449,187	70,087,484	96.76%	64,428,135	64,428,135	2.2	1.25	1.43%	3.20%	28.15	1.44	0.63	0.08	0.50
Alive	2760	6	2	3	F	6	WGS	100	17,553,427	75,740,025	95.45%	62,562,554	62,562,554	2.2	1.41	1.02%	2.85%	24.26	2.40	0.77	-	-
Alive	2760	6	2	2	G	2	WGS	100	18,233,453	79,627,508	98.14%	69,294,135	69,294,135	2.4	1.54	23.12%	2.85%	18.22	1.55	0.84	0.02	0.57
Alive	2760	6	13	3	C	3	WGS	100	45,654,281	16,942,152	94.59%	91,414%	91,414%	1.3	1.21	0.24%	2.72%	18.16	-1.07	0.47	-	-
Alive	2760	6	8	3	F	6	WGS	100	34,738,187	68,651,306	94.47%	91,414%	91,414%	1.3	1.21	0.24%	2.72%	18.16	-1.07	0.47	-	-
Alive	2760	6	11	3	H	6	WGS	100	32,879,120	65,336,340	94.50%	91,414%	91,414%	1.3	1.21	0.24%	2.72%	18.16	-1.07	0.47	-	-
Deceased	2760	6	3	2	G	6	WGS	100	42,094,130	44,118,480	94.00%	91,414%	91,414%	1.3	1.21	0.24%	2.72%	18.16	-1.07	0.47	-	-
Alive	2760	6	16	2	D	4	WGS	100	17,902,132	70,044,184	94.00%	64,428,135	64,428,135	2.1	1.35	0.83%	1.59%	13.71	-0.44	0.12	0.04	0.46
Alive	2760	6	4	2	C	2	WGS	100	43,827,133	70,748,658	94.00%	64,428,135	64,428,135	2.1	1.35	0.83%	1.59%	13.71	-0.44	0.12	0.04	0.46
Deceased	2760	6	10	3	F	6	WGS	100	50,888,182	84,517,152	93.12%	68,794%	68,794%	1.7	1.47	0.89%	2.28%	21.82	-1.18	0.61	1.00	0.51
Deceased	2760	6	3	2	G	6	WGS	100	55,343,363	77,027,119	95.00%	92,894%	92,894%	2.1	1.30	34.00%	2.33%	20.49	2.08	1.00	0.88	0.82
Deceased	2760	6	4	2	G	6	WGS	100	46,744,147	93,440,484	94.44%	93,776%	93,776%	2.6	1.57	1.24%	2.33%	14.24	-1.48	0.40	0.34	0.64
Deceased	2760	6	17	4	D	6	WGS	100	42,653,441	85,387,182	92.35%	88,474%	88,474%	2.1	1.60	7.20%	2.40%	18.06	-1.02	0.47	-	-
Deceased	2760	6	13	3	F	2	WGS	100	58,899,144	77,718,602	94.55%	93,174%	93,174%	2.1	1.47	1.20%	1.54%	14.12	-1.97	0.31	-	-
Alive	2760	6	15	1	A	4	WGS	100	54,501,147	89,116,684	96.00%	93,894%	93,894%	2.0	1.45	0.74%	2.16%	17.66	1.03	0.76	-	-
Alive	2760	6	3	1	D	2	WGS	100	57,941,155	78,127,110	93.66%	96,694%	96,694%	2.1	1.31	2.16%	2.30%	13.18	-0.97	0.56	-	-
Deceased	2760	6	16	4	A	6	WGS	100	44,154,159	85,302,115	95.15%	93,776%	93,776%	1.6	1.21	0.78%	2.40%	17.45	-1.37	0.43	0.43	0.43
Deceased	2760	6	2	2	C	4	WGS	100	51,488,178	84,517,152	95.45%	93,776%	93,776%	2.0	1.45	2.02%	2.32%	18.75	0.78	0.71	-	-
Deceased	2760	6	13	3	C	4	WGS	100	51,921,187	105,942,784	94.23%	96,914%	96,914%	2.1	1.38	41.49%	7.40%	20.65	17.38	1.00	1.00	1.00
Deceased	2760	6	11	3	H	5	WGS	100	38,469,135	78,987,488	94.04%	91,009%	91,009%	2.1	1.35	5.10%	1.05%	13.47	1.41	0.82	1.00	0.91
Alive	2760	6	3	2	G	1	WGS	100	40,344,166	80,640,082	96.73%	94,004%	94,004%	2.4	1.45	0.89%	2.07%	14.39	2.02	0.46	-	-
Alive	2760	6	7	3	F	6	WGS	100	36,654,456	71,508,265	95.30%	93,076%	93,076%	2.3	1.35	0.85%	3.00%	13.37	-1.38	0.40	0.01	0.21
Deceased	2760	6	14	2	A	4	WGS	100	34,053,153	68,106,308	95.28%	93,776%	93,776%	2.0	1.32	0.78%	1.62%	14.07	-1.68	0.43	-	-
Alive	2760	6	17	2	B	1	WGS	100	40,808,468	81,212,196	95.07%	91,474%	91,474%	2.1	1.27	0.78%	2.40%	14.52	-1.10	0.50	0.00	0.15
Deceased	2760	6	8	3	F	6	WGS	100	47,745,133	83,433,078	95.20%	91,474%	91,474%	2.3	1.26	0.69%	3.80%	13.10	-1.23	0.50	0.15	0.32
Deceased	2760	6	4	2	C	6	WGS	100	42,553,109	85,106,016	95.44%	94,118%	94,118%	2.6	1.17	0.82%	3.90%	27.06	-1.26	0.44	-	-
Deceased	2760	6	1	2	C	4	WGS	100	38,790,166	77,581,171	95.33%	94,404%	94,404%	2.3	1.27	0.83%	4.80%	28.44	-0.08	0.82	-	-
Alive	2760	6	2	2	C	7	WGS	100	34,873,135	69,646,050	95.17%	93,476%	93,476%	2.1	1.27	0.87%	2.40%	14.35	1.11	0.73	-	-
Alive	2760	6	3	2	G	1	WGS	100	40,594,009	81,012,000	96.79%	94,414%	94,414%	2.3	1.20	0.69%	2.67%	14.75	-0.37	0.43	0.00	0.19
Alive	2760	6	4	2	G	1	WGS	100	31,708,127	64,133,254	96.79%	93,904%	93,904%	1.9	1.18	1.70%	1.75%	14.75	-1.11	0.61	-	-
Deceased	2760	6	16	1	C	3	WGS	100	40,308,026	80,731,052	94.29%	91,224%	91,224%	2.3	1.23	0.69%	2.75%	15.36	-1.84	0.55	-	-
Alive	2760	6	5	2	G	2	WGS	100	40,517,450	81,044,100	94.87%	92,894%	92,894%	2.1	1.46	0.84%	2.80%	13.60	2.13	0.84	-	-
Deceased	2760	6	4	2	C	5	WGS	100	36,176,457	72,342,112	92.51%	88,178%	88,178%	2.1	1.40	0.80%	2.80%	13.12	2.13	0.79	-	-
Alive	2760	6	5	2	G	2	WGS	100	48,904,142	97,808,484	94.59%	92,894%	92,894%	2.6	1.77	0.70%	2.76%	13.39	-0.63	0.64	0.00	0.04
Alive	2760	6	2	2	G	5	WGS	100	57,102,127	114,384,966	93.47%	98,144%	98,144%	2.1	1.23	0.24%	2.80%	27.32	-0.67	0.64	0.00	0.04
Deceased	2760	6	11	3	H	4	WGS	100	36,397,133	73,187,478	94.40%	91,494%	91,494%	2.0	1.36	1.70%	1.41%	13.13	-0.63	0.50	-	-
Deceased	2760	6	9	3	F	4	WGS	100	35,397,147	71,185,094	93.63%	96,934%	96,934%	2.1	1.36	1.09%	1.60%	14.44	1.29	0.91	-	-
Deceased	2760	6	10	3	F	4	WGS	100	36,648,169	70,719,914	95.70%	91,959%	91,959%	2.3	1.37	0.70%	2.47%	16.15	-1.61	0.25	0.50	0.37
Deceased	2760	6	14	1	A	4	WGS	100	45,758,133	80,535,788	94.93%	91,959%	91,959%	2.7	1.53	1.59%	2.47%	16.15	-1.61	0.25	0.50	0.37
Deceased	2760	6	13	3	F	3	WGS	100	38,112,101	76,225,402	95.59%	91,959%	91,959%	2.1	1.44	1.44%	2.80%	14.40	-1.72	0.31	0.00	0.18
Deceased	2760	6	7	2	C	3	WGS	100	40,113,113	80,246,218	94.29%	90,774%	90,774%	2.3	1.27	0.64%	2.80%	14.40	-1.72	0.31	0.00	0.18
Deceased	2760	6	11	3	H	4	WGS	100	38,448,181	72,138,182	95.44%	91,959%	91,959%	2.0	1.23	0.80%	2.80%	13.97	2.01	0.78	-	-
Deceased	2760	6	11	3	H	4	WGS	100	37,054,451	74,188,102	95.76%	91,959%	91,959%	1.9	1.42	0.74%	1.40%	13.35	1.18	0.78	-	-
Deceased	2760	6	14	1	A	4	WGS	100	36,908,087	73,185,970	96.24%	91,959%	91,959%	1.9	1.42	1.09%	1.60%	13.15	-0.63	0.50	-	-
Deceased	2760	6	10	3	F	4	WGS	100	45,758,133	80,535,788	94.93%	91,959%	91,959%	2.7	1.53	1.59%	2.47%	16.15	-1.61	0.25	0.50	0.37
Deceased	2760	6	10	3	F	4	WGS	100	36,648,169	70,719,914	95.70%	91,959%	91,959%	2.3	1.37	0.70%	2.47%	16.15	-1.61	0.25	0.50	0.37
Deceased	2760	6	14	1	A	4	WGS	100	45,758,133	80,535,788	94.93%	91,959%	91,959%	2.7	1.53	1.59%	2.47%	16.15	-1.61	0.25	0.50	0.37
Deceased	2760	6	10	3	F	4	WGS	100	45,758,133	80,535,788	94.93%	91,959%	91,959%	2.7	1.53	1.59%	2.47%	16.15	-1.61	0.25	0.50	0.37
Deceased	2760	6	10	3	F	4	WGS	100	45,758,133	80,535,788	94.93%	91,959%	91,959%	2.7	1.53	1.59%	2.47%	16.15	-1.61	0.25	0.50	0.37
Deceased	2760	6	10	3	F	4	WGS	100	45,758,133	80,535,788	94.93%	91,959%	91,959%	2.7	1.53	1.59%	2.47%	16.15	-1.61	0.25	0.50	0.37
Deceased	2760	6	10	3	F	4	WGS	100	45,758,133	80,535,788	94.93%	91,959%	91,959%	2.7	1.53	1.59%	2.47%	16.15	-1.61	0.25	0.50	0.37
Deceased	2760	6	10	3	F	4	WGS	100	45,758,133	80,535,788	94.93%	91,959%	91,959%	2.7	1.53	1.59%	2.47%	16.15	-1.61	0.25	0.50	0.37
Deceased	2760	6	10	3	F	4	WGS	100	45,758,133	80,535,788	94.93%	91,959%	91,959%	2.7	1.53	1.59%	2.47%	16.15	-1.61	0.25	0.50	0.37
Deceased	2760	6	10	3	F	4	WGS	100	45,758,133	80,535,788	94.93%	91,959%	91,959%	2.7	1.53	1.59%	2.47%	16.15	-1.61	0.25	0.50	0.37
Deceased	2760	6	10	3	F	4	WGS	100	45,758,133	80,535,788	94.93%	91,959%	91,959%	2.7	1.53	1.59%	2.47%	16.15	-1.61	0.25	0.50	0.37
Deceased	2760	6	10	3	F	4	WGS	100	45,758,133	80,535,788												

[illegible]

Survival status	Days since birth	eDNA extraction batch	Genomic library batch	Sequencing machine	Flow cell	Lane of flow cell	Sequencing approach	Read type	Read length	Number of sequenced fragments	Number of sequenced reads	Percent of 0.28 bases	Percent of 0.39 bases	Reads mapped to genome	Whole genome coverage	Estimated 5'-end fraction	Whole-genome fraction	Single molecule C.A. frequency changes	Single molecule C.A. frequency (per Mb of testable bases)	Regional difference in single molecule C.A. frequency (per Mb of testable bases)	GENOME score _{20k}	DETA score	GENOME / DETA score
Alive	2754	13	8	2	1	3	WGS	Paired-end	100	40,569,375	61,796,348	34.74%	52.44%	34,892,138	2.6	1.27	31.66%	3,244	29.17	-2.36	0.72	-	-
Deceased	2754	14	23	1	4	7	WGS	Paired-end	100	37,575,568	75,154,906	34.77%	52.71%	61,520,748	2.6	2.23	31.66%	3,297	18.19	-1.91	0.75	-	-
Deceased	2754	15	8	3	3	3	WGS	Paired-end	100	45,529,411	82,246,842	35.25%	52.61%	82,394,103	2.7	1.34	31.66%	3,996	11.28	-0.71	0.71	-	-
Deceased	2844	15	23	3	3	3	WGS	Paired-end	100	48,681,144	86,327,188	34.85%	49.13%	84,594,644	2.7	1.35	0.73%	1,795	14.41	-0.06	0.12	0.06	0.06
Alive	2752	15	2	7	6	6	WGS	Paired-end	100	32,763,141	65,526,322	36.46%	54.57%	64,117,341	2.6	1.44	0.73%	4,732	14.46	-0.46	0.32	-	-
Alive	2752	15	1	3	7	2	WGS	Paired-end	100	22,584,745	45,169,470	35.43%	49.47%	38,338,133	1.1	1.38	1.07%	1,039	13.69	-0.46	0.39	-	-
Alive	2752	15	28	3	7	2	WGS	Paired-end	100	33,844,131	67,683,647	36.53%	49.29%	61,538,134	2.6	1.39	0.73%	4,810	10.74	5.45	0.39	0.15	0.38
Alive	2769	15	16	1	6	3	WGS	Paired-end	100	42,892,382	85,784,764	36.02%	53.37%	79,481,165	2.6	1.78	0.73%	3,904	13.76	-1.37	0.14	-	-
Deceased	514	15	23	3	6	3	WGS	Paired-end	100	25,567,964	71,173,948	39.42%	52.67%	70,496,407	2.4	1.42	4.10%	2,696	19.48	-1.91	0.31	-	-
Deceased	757	15	3	1	1	2	WGS	Paired-end	100	28,708,079	58,416,158	38.10%	49.11%	53,145,124	1.8	1.32	1.07%	1,313	29.84	0.39	0.62	-	-
Deceased	48	15	8	3	7	6	WGS	Paired-end	100	17,154,344	74,702,764	36.12%	49.48%	69,340,127	2.4	1.24	1.07%	1,794	21.45	3.03	0.92	-	-
Alive	2767	15	23	3	7	6	WGS	Paired-end	100	25,598,124	58,416,148	38.11%	49.37%	47,328,135	1.1	1.14	0.73%	1,513	20.63	7.12	0.37	0.09	0.46
Deceased	483	15	10	3	6	7	WGS	Paired-end	100	35,321,322	70,642,644	42.15%	49.30%	38,036,111	1.9	1.24	0.89%	1,141	33.27	-1.45	0.38	0.68	0.68
Alive	2764	15	8	1	6	1	WGS	Paired-end	100	32,243,172	76,482,344	36.93%	50.76%	69,443,187	2.2	1.68	0.73%	1,901	14.34	-0.39	0.17	0.16	0.34
Alive	2764	15	7	3	6	8	WGS	Paired-end	100	44,428,694	88,857,388	36.83%	53.89%	88,539,165	2.6	1.28	0.44%	1,583	21.27	-0.11	0.59	0.95	0.78
Alive	2764	15	4	2	6	3	WGS	Paired-end	100	33,124,953	66,249,906	36.85%	52.47%	62,590,178	2.9	1.36	0.89%	1,459	34.56	-2.44	0.50	-	-
Alive	2764	15	24	3	7	7	WGS	Paired-end	100	53,770,737	107,541,484	34.13%	50.29%	94,697,133	3.2	1.40	1.75%	1,984	27.15	-1.09	0.46	0.11	0.28
Alive	2764	15	19	1	6	6	WGS	Paired-end	100	49,449,151	98,898,302	34.89%	52.29%	74,693,346	2.3	1.42	9.61%	2,409	19.26	-4.74	0.09	0.17	0.42
Alive	2763	16	22	3	6	3	WGS	Paired-end	100	50,343,603	100,687,206	35.70%	52.67%	84,973,166	3.0	1.61	3.76%	1,970	19.66	-1.70	0.17	-	-
Deceased	567	16	5	2	6	7	WGS	Paired-end	100	49,941,124	99,882,248	35.77%	51.77%	91,283,166	2.7	1.64	0.49%	1,107	13.02	-0.55	0.45	-	-
Deceased	184	16	12	3	6	1	WGS	Paired-end	100	39,412,084	78,824,168	38.24%	52.95%	69,970,108	2.3	1.74	2.80%	1,866	14.51	-0.88	0.33	0.68	0.71
Alive	2763	16	6	2	1	3	WGS	Paired-end	100	36,463,989	72,927,978	38.04%	50.34%	67,970,108	2.1	1.55	1.85%	2,775	35.52	-0.39	0.36	-	-
Alive	2763	16	12	3	1	5	WGS	Paired-end	100	64,539,196	129,078,112	35.77%	52.12%	98,693,114	5.7	1.09	4.06%	2,266	20.69	-0.09	0.59	-	-
Deceased	35	16	20	1	1	3	WGS	Paired-end	100	48,254,878	96,509,756	39.14%	50.61%	76,260,173	2.5	1.57	1.97%	1,136	19.68	2.15	0.87	1.00	0.93
Alive	2762	16	10	1	6	4	WGS	Paired-end	100	35,442,394	70,884,788	39.29%	50.75%	67,246,132	2.2	2.09	0.80%	1,364	19.08	-3.18	0.19	0.30	0.10
Alive	2762	16	3	2	6	7	WGS	Paired-end	100	47,628,715	95,257,438	37.15%	49.15%	82,754,119	2.6	1.44	0.88%	1,666	15.59	-3.01	0.33	0.30	0.16
Alive	2762	16	2	1	7	3	WGS	Paired-end	100	45,488,329	90,976,778	36.60%	49.15%	82,254,177	2.6	1.32	0.52%	1,361	11.62	-1.36	0.36	-	-
Deceased	42	16	12	1	6	4	WGS	Paired-end	100	48,337,715	96,675,438	37.49%	49.19%	82,025,190	2.6	1.47	4.80%	2,122	19.43	1.34	0.39	1.00	0.39
Deceased	2761	16	15	1	4	8	WGS	Paired-end	100	47,468,828	94,937,656	34.65%	49.17%	73,794,173	2.4	1.56	2.81%	3,947	20.97	4.55	0.97	1.00	0.96
Alive	2761	16	3	1	6	6	WGS	Paired-end	100	37,471,341	74,942,702	39.44%	49.15%	68,725,169	2.3	1.68	0.70%	1,520	34.78	-0.46	0.53	-	-
Deceased	1060	16	4	2	6	3	WGS	Paired-end	100	37,886,140	75,778,280	36.73%	50.15%	70,538,164	2.3	1.57	0.46%	1,917	14.28	-0.05	0.82	-	-
Alive	2760	16	5	2	6	7	WGS	Paired-end	100	49,398,194	98,796,388	35.43%	49.29%	84,042,198	2.8	2.20	0.89%	2,452	13.14	1.06	0.77	0.35	0.65
Deceased	2762	16	6	2	1	5	WGS	Paired-end	100	49,449,124	98,898,248	35.03%	49.29%	74,622,115	2.8	1.24	0.73%	1,116	29.44	-1.57	0.15	-	-
Alive	2687	17	23	3	6	3	WGS	Paired-end	100	37,828,145	75,656,290	32.12%	49.16%	68,076,182	1.6	1.51	0.90%	1,686	13.45	1.36	0.48	0.38	0.70
Deceased	2182	17	11	2	6	4	WGS	Paired-end	100	48,394,080	96,788,160	32.12%	47.71%	82,122,115	2.4	1.38	0.80%	2,489	14.15	-1.36	0.18	0.34	0.70
Deceased	2761	17	23	3	6	3	WGS	Paired-end	100	45,754,128	91,508,256	38.25%	49.17%	77,551,143	2.6	1.36	0.73%	3,237	21.68	1.38	0.74	-	-
Alive	2687	17	14	1	6	4	WGS	Paired-end	100	45,075,601	90,151,202	38.19%	49.14%	82,047,147	2.6	1.54	0.70%	1,902	14.76	-0.45	0.86	-	-
Deceased	2761	17	15	1	6	3	WGS	Paired-end	100	37,794,140	75,588,280	32.01%	48.77%	68,124,111	2.2	1.28	0.80%	1,839	14.15	-1.36	0.18	0.34	0.70
Deceased	2761	17	23	3	6	3	WGS	Paired-end	100	47,828,145	95,656,290	32.12%	49.16%	74,622,115	2.4	1.38	0.80%	2,489	14.15	-1.36	0.18	0.34	0.70
Deceased	36	17	11	2	6	1	WGS	Paired-end	100	37,794,140	75,588,280	32.01%	48.77%	68,124,111	1.9	1.51	0.44%	1,412	14.28	-0.48	0.60	0.27	0.73
Deceased	2761	17	10	3	6	1	WGS	Paired-end	100	49,154,180	98,308,360	32.01%	47.71%	79,620,115	2.6	1.17	0.93%	1,539	33.54	0.41	0.71	1.06	0.58
Alive	2684	17	11	3	6	3	WGS	Paired-end	100	45,428,634	90,857,268	36.88%	49.94%	81,508,166	2.6	1.58	0.74%	2,619	34.17	-4.34	0.69	0.65	0.38
Deceased	2761	17	15	1	6	3	WGS	Paired-end	100	37,794,140	75,588,280	32.01%	48.77%	68,124,111	2.2	1.29	1.06%	1,412	14.28	-0.48	0.60	0.27	0.73
Deceased	2761	17	15	1	6	3	WGS	Paired-end	100	37,794,140	75,588,280	32.01%	48.77%	68,124,111	2.2	1.29	1.06%	1,412	14.28	-0.48	0.60	0.27	0.73
Deceased	360	17	14	1	6	1	WGS	Paired-end	100	45,954,675	91,909,352	36.30%	49.24%	80,735,185	2.2	1.31	0.73%	2,667	16.79	1.21	0.62	1.00	0.66
Deceased	2760	17	15	1	6	7	WGS	Paired-end	100	30,454,173	60,908,346	34.80%	52.49%	51,831,146	1.7	1.31	0.80%	1,133	11.86	-1.15	0.45	0.62	0.43
Alive	2698	17	20	2	6	3	WGS	Paired-end	100	41,514,149	83,028,298	35.97%	53.07%	77,123,119	2.3	1.39	0.46%	1,107	17.82	1.31	0.91	-	-
Deceased	2760	17	21	3	6	3	WGS	Paired-end	100	35,668,181	71,336,362	35.59%	51.51%	69,443,169	1.8	1.34	0.46%	1,631	19.83	-0.30	0.58	-	-
Deceased	56	17	20	2	6	1	WGS	Paired-end	100	41,833,190	83,666,380	35.94%	50.44%	74,697,171	2.6	1.37	3.78%	2,637	15.77	0.25	0.97	1.00	0.93
Deceased	2761	17	8	3	6	3	WGS	Paired-end	100	37,471,341	74,942,702	39.44%	49.15%	68,725,169	2.2	1.51	0.84%	1,294	17.12	-0.49	0.36	-	-
Alive	2683	18	10	1	4	7	WGS	Paired-end	100	35,442,115	70,884,238	39.29%	50.75%	67,246,132	2.3	1.66	0.46%	1,075	18.08	-0.35	0.57	-	-
Deceased	72	18	23	2	8	3	WGS	Paired-end	100	47,844,406	95,688,816	36.45%	49.18%	79,691,112	2.5	1.35	4.43%	1,874	15.99	-1.06	0.47	1.00	0.73
Deceased	2684	18	23	3	6	3	WGS	Paired-end	100	45,661,171	91,322,342	35.94%	49.94%	80,735,185	2.6	1.35	0.51%	1,896	14.28	-0.46	0.69	0.39	0.12

[illegible]

Survival status	Days since adjuv	CRNA extraction method	Sequencing machine	Flow cell	Library batch	Sequencing approach	Read type	Read length	Number of sequenced fragments	Number of sequenced reads	Percent of Q30 bases	Percent of Q30 bases	Reads mapped to genome	Whole genome coverage	Estimated coverage	Inter-Mi fusion	Single molecule C.A frequency of changes	Single molecule C.A frequency per 10k of reads	Single molecule C.A frequency per 10k of reads	Regional difference in single molecule C.A frequency per 10k of reads	CRNA score	CRNA score
Alive	2641	22	12	1	1	WGS	Paired-end	100	39,711,871	76,423,742	95.54%	95.54%	72,654,308	7.3	2.37	0.15%	2,444	17.32	17.32	1.55	0.16	-
Alive	2641	22	12	1	1	WGS	Paired-end	100	65,613,121	95,354	95.55%	95.55%	76,594,431	7.6	1.33	0.15%	6,519	40.97	40.97	2.12	0.17	-
Alive	2642	23	13	1	1	WGS	Paired-end	100	40,431,934	96,901,182	95.54%	95.54%	76,559,458	7.4	1.39	0.14%	2,181	13.96	13.96	0.31	0.15	-
Deceased	2643	23	13	1	1	WGS	Paired-end	100	65,520,751	96,600,746	95.53%	95.46%	76,433,945	7.5	1.39	0.17%	2,577	18.75	18.75	0.31	0.16	0.87
Alive	2639	22	12	1	1	WGS	Paired-end	100	37,496,425	75,899,549	95.43%	95.43%	68,232,317	7.2	1.35	0.51%	2,619	18.27	18.27	0.32	0.15	0.07
Alive	2638	22	12	1	1	WGS	Paired-end	100	40,168,260	68,738,760	95.73%	95.46%	74,021,259	7.4	1.37	0.40%	3,476	22.46	22.46	0.25	0.09	0.12
Alive	2638	22	12	1	1	WGS	Paired-end	100	31,611,641	87,235,462	94.94%	95.13%	67,624,977	7.3	1.38	0.38%	1,468	14.11	14.11	0.35	0.07	0.19
Deceased	264	21	10	1	1	WGS	Paired-end	100	42,217,351	84,441,506	94.64%	92.90%	77,553,616	7.5	1.10	0.24%	2,580	22.60	22.60	0.32	0.12	-
Deceased	264	22	11	1	1	WGS	Paired-end	100	42,443,851	84,451,570	95.35%	94.70%	77,897,428	7.5	1.19	0.24%	2,787	23.92	23.92	0.34	0.09	0.37
Alive	2634	21	10	1	1	WGS	Paired-end	100	35,100,831	70,617,162	95.24%	95.75%	66,151,608	7.1	1.34	0.51%	1,880	12.90	12.90	0.79	1.00	0.90
Alive	2634	21	10	1	1	WGS	Paired-end	100	26,718,841	57,394,431	94.64%	94.64%	51,734,431	6.7	1.33	0.41%	2,087	19.49	19.49	0.37	0.34	-
Deceased	651	22	12	1	1	WGS	Paired-end	100	36,193,711	71,785,142	92.03%	92.70%	58,688,054	6.9	1.67	1.13%	1,775	11.41	11.41	0.77	0.73	-
Alive	2632	21	10	1	1	WGS	Paired-end	100	40,917,512	81,834,034	94.72%	92.93%	74,693,693	7.4	1.29	0.34%	1,958	13.82	13.82	0.45	0.09	0.36
Alive	2632	21	10	1	1	WGS	Paired-end	100	31,100,382	103,361,764	94.03%	94.69%	102,019,682	10.3	1.35	1.40%	3,430	22.47	22.47	0.68	0.12	0.37
Alive	2632	21	10	1	1	WGS	Paired-end	100	24,276,664	34,071,828	94.59%	92.24%	72,860,064	7.5	1.17	0.37%	2,083	21.77	21.77	0.71	0.06	-
Deceased	71	23	13	1	1	WGS	Paired-end	100	34,025,602	66,251,804	94.69%	93.23%	58,638,068	7.4	1.31	0.37%	1,767	11.77	11.77	1.29	0.62	1.00
Alive	2632	23	13	1	1	WGS	Paired-end	100	47,066,602	94,138,804	95.63%	93.32%	76,833,536	10.3	1.31	1.00%	2,317	14.39	14.39	0.73	0.24	0.09
Deceased	2100	23	13	1	1	WGS	Paired-end	100	27,409,654	44,217,712	95.73%	93.07%	41,873,662	7.3	1.19	0.40%	1,036	11.74	11.74	0.66	0.71	-
Deceased	78	23	13	1	1	WGS	Paired-end	100	45,794,361	97,480,738	94.67%	94.84%	77,916,074	10.3	1.21	0.43%	3,530	23.81	23.81	0.81	0.33	-
Alive	2631	23	13	1	1	WGS	Paired-end	100	41,401,397	82,814,794	96.15%	95.46%	77,056,720	7.5	1.31	0.13%	6,234	40.54	40.54	0.83	0.33	-
Alive	2630	23	13	1	1	WGS	Paired-end	100	41,944,824	88,989,262	96.07%	95.76%	86,551,768	7.7	1.31	0.09%	2,037	12.71	12.71	1.21	0.33	-
Alive	2630	23	13	1	1	WGS	Paired-end	100	47,732,934	85,261,128	98.29%	94.23%	77,827,728	10.3	1.39	0.47%	2,174	14.20	14.20	0.83	0.33	-
Deceased	42	23	13	1	1	WGS	Paired-end	100	37,268,197	74,531,214	98.57%	94.78%	68,512,402	7.2	1.39	2.40%	1,588	13.49	13.49	0.81	0.33	-
Deceased	282	23	13	1	1	WGS	Paired-end	100	45,168,910	90,889,940	95.67%	95.05%	76,533,453	10.3	1.36	0.13%	7,332	45.96	45.96	1.31	0.84	1.00
Alive	2617	23	13	1	1	WGS	Paired-end	100	39,246,397	77,492,584	95.92%	95.92%	66,738,495	7.2	1.44	0.13%	3,974	32.97	32.97	0.62	0.46	0.43
Deceased	3165	25	15	1	1	WGS	Paired-end	100	39,956,841	79,911,621	95.63%	92.64%	75,475,739	7.4	1.34	0.61%	1,360	11.43	11.43	0.74	0.24	-
Alive	2617	23	13	1	1	WGS	Paired-end	100	43,714,604	87,961,603	95.63%	92.64%	76,419,471	7.6	1.26	0.37%	3,708	18.19	18.19	0.67	0.36	-
Alive	2617	23	13	1	1	WGS	Paired-end	100	39,342,615	87,145,980	96.03%	92.38%	61,883,081	7.0	1.48	0.20%	2,786	33.15	33.15	1.29	0.96	-
Alive	2618	23	13	1	1	WGS	Paired-end	100	37,715,008	85,428,616	96.42%	93.24%	61,231,667	7.0	1.27	0.38%	2,455	13.82	13.82	0.32	0.24	-
Alive	2618	23	13	1	1	WGS	Paired-end	100	47,636,056	95,251,912	95.47%	92.78%	76,927,397	10.3	1.44	0.04%	2,338	14.75	14.75	0.68	0.71	-
Deceased	344	24	14	1	1	WGS	Paired-end	100	34,991,941	69,032,602	95.63%	92.64%	66,411,116	7.1	1.31	0.39%	1,930	18.24	18.24	1.19	0.84	0.46
Deceased	201	24	14	1	1	WGS	Paired-end	100	42,036,904	84,073,806	95.46%	92.56%	77,861,139	7.5	1.37	0.47%	2,442	14.27	14.27	0.81	0.81	0.41
Alive	2615	24	14	1	1	WGS	Paired-end	100	35,784,930	77,432,000	95.73%	95.11%	66,233,472	7.3	1.39	0.13%	3,559	24.61	24.61	0.37	0.37	0.27
Alive	2615	24	14	1	1	WGS	Paired-end	100	41,437,762	84,846	94.67%	93.47%	78,827,432	10.3	1.39	0.08%	2,545	14.60	14.60	0.89	0.48	-
Alive	2615	24	14	1	1	WGS	Paired-end	100	45,332,563	97,245,186	95.63%	93.75%	69,833,046	9.9	1.25	0.31%	3,345	14.67	14.67	0.78	0.74	0.47
Alive	2615	24	14	1	1	WGS	Paired-end	100	29,592,112	66,730,424	95.29%	92.66%	55,939,252	7.0	1.04	0.36%	2,239	17.92	17.92	0.26	0.34	-
Deceased	2421	24	14	1	1	WGS	Paired-end	100	39,483,147	78,920,094	95.24%	94.24%	72,666,253	7.6	1.49	0.08%	3,082	22.60	22.60	0.56	0.53	-
Alive	2613	24	14	1	1	WGS	Paired-end	100	42,633,716	83,033,152	94.78%	92.44%	77,184,132	7.3	1.61	0.17%	2,789	18.27	18.27	0.81	0.81	-
Alive	2613	24	14	1	1	WGS	Paired-end	100	41,736,004	87,489,466	94.20%	90.76%	76,904,538	7.3	1.37	0.13%	2,679	18.95	18.95	1.13	0.48	0.42
Alive	2613	24	14	1	1	WGS	Paired-end	100	45,070,481	90,140,966	94.67%	93.88%	83,688,267	7.7	1.33	0.10%	2,864	14.71	14.71	0.43	0.36	-
Alive	2613	24	14	1	1	WGS	Paired-end	100	42,168,576	84,416,752	96.03%	93.47%	77,868,133	7.6	1.32	0.73%	3,384	15.41	15.41	0.33	0.38	-
Alive	2610	24	14	1	1	WGS	Paired-end	100	39,867,723	86,067,436	95.75%	90.11%	76,099,397	7.3	1.31	0.31%	4,078	18.19	18.19	0.60	0.05	0.16
Alive	2610	24	14	1	1	WGS	Paired-end	100	50,864,444	101,727,068	95.18%	92.75%	74,630,989	7.7	1.45	0.47%	3,224	15.42	15.42	0.67	0.36	-
Deceased	277	24	14	1	1	WGS	Paired-end	100	42,162,656	84,320,152	95.64%	92.69%	72,889,469	7.5	1.20	0.16%	3,663	25.94	25.94	0.69	0.72	0.95
Alive	2613	25	15	1	1	WGS	Paired-end	100	39,067,134	74,013,388	95.63%	90.73%	61,483,617	7.1	1.33	0.31%	1,482	13.89	13.89	0.78	0.21	-
Alive	2613	25	15	1	1	WGS	Paired-end	100	45,540,334	94,268,666	95.63%	94.06%	82,669,174	9.7	1.27	0.77%	2,450	16.89	16.89	0.43	0.47	0.51
Alive	2617	25	15	1	1	WGS	Paired-end	100	40,949,483	81,886,176	97.03%	94.82%	76,714,073	7.5	1.26	0.34%	2,211	14.17	14.17	0.32	0.76	-
Alive	2617	25	15	1	1	WGS	Paired-end	100	77,134,013	143,030,038	93.98%	90.41%	114,738,668	14.4	1.04	0.09%	1,108	13.84	13.84	1.21	0.43	-
Alive	2617	25	15	1	1	WGS	Paired-end	100	41,607,064	88,413,912	96.66%	91.06%	81,009,131	7.7	1.21	0.30%	1,906	14.57	14.57	0.19	0.19	-
Deceased	266	25	15	1	1	WGS	Paired-end	100	35,681,105	71,371,370	94.27%	91.67%	61,558,948	7.0	1.34	0.48%	2,058	12.75	12.75	0.13	0.13	0.69
Deceased	261	25	15	1	1	WGS	Paired-end	100	42,899,246	85,812,912	95.95%	94.13%	76,142,518	7.6	1.32	0.08%	2,982	20.91	20.91	0.66	0.53	-
Deceased	261	25	15	1	1	WGS	Paired-end	100	42,733,024	81,544,166	95.36%	95.21%	82,134,437	7.7	1.13	1.01%	6,575	41.15	41.15	0.96	0.67	0.31
Deceased	34	25	15	1	1	WGS	Paired-end	100	42,836,394	93,346,966	96.57%	92.64%	78,600,159	7.6	1.68	0.43%	3,170	14.14	14.14	0.83	0.21	-

Screened status	Opt. drive	cd/ha extraction batch	Economic library fraction	Sequencing machine	Flow cell	Time of flow cell	Sequencing approach	Read length	Number of sequenced fragments	Number of sequenced reads	Percent of 120 bases	Percent of 150 bases	Reads mapped to genome	Whole genome coverage	Estimated 5-seq fraction	Sequencing C.A. frequency (per Mb of sequenced bases)	Sequencing C.A. frequency (per Mb of available bases)	Regional effective coverage (per Mb of available bases)	QEMM score	QEMM score	QEMM score	
Alive	2611	25	2	2	2	2	WGS	100	41,131,100	21,241,000	94.03%	94.03%	75,741,100	2.4	1.26	0.75%	2.17%	18.11	-1.24	0.51	0.09	0.15
Alive	2612	25	21	2	2	2	WGS	100	27,942,130	75,380,270	93.89%	93.89%	58,589,870	2.2	1.32	1.37%	2.68%	25.68	-3.45	0.17	-	-
Deceased	2613	25	2	2	2	2	WGS	100	31,681,142	71,253,221	96.29%	96.29%	61,783,100	2.1	1.18	0.93%	1.93%	18.12	-4.34	0.03	-	-
Alive	2614	25	12	2	2	2	WGS	100	46,039,244	91,039,145	94.13%	94.13%	61,031,204	2.8	1.41	0.83%	2.48%	15.16	-2.16	0.23	0.08	0.17
Alive	2615	25	12	2	2	2	WGS	100	46,157,122	92,175,044	95.31%	95.31%	74,411,641	2.4	1.24	0.85%	3.24%	14.61	-3.16	0.19	-	-
Deceased	2616	25	12	2	2	2	WGS	100	54,012,608	108,023,700	95.96%	95.96%	99,981,537	3.1	1.30	0.94%	4.54%	19.66	-3.46	0.49	1.06	1.00
Alive	2617	25	17	2	2	2	WGS	100	39,325,157	78,651,714	94.86%	94.86%	79,929,684	2.2	1.18	0.73%	2.74%	15.56	-1.82	0.35	0.04	0.11
Alive	2618	25	12	2	2	2	WGS	100	45,554,544	91,109,088	93.48%	93.48%	75,821,439	2.8	1.31	2.15%	2.84%	18.03	-1.68	0.24	-	-
Alive	2619	25	14	2	2	2	WGS	100	36,329,161	72,658,712	93.61%	93.61%	68,042,137	2.1	1.30	0.88%	1.93%	18.17	-0.32	0.43	-	-
Deceased	2620	25	13	2	2	2	WGS	100	34,138,933	107,481,346	94.91%	94.91%	98,032,017	3.0	1.31	1.13%	3.13%	18.58	-1.39	0.42	0.80	0.81
Deceased	2621	25	13	2	2	2	WGS	100	48,173,975	96,347,950	94.79%	94.79%	94,138,728	2.7	1.30	1.83%	2.48%	14.91	-3.14	0.19	-	-
Alive	2622	25	15	2	2	2	WGS	100	31,110,121	62,220,242	92.44%	92.44%	62,135,100	2.7	1.18	0.96%	1.13%	14.01	-0.32	0.59	-	-
Alive	2623	25	20	2	2	2	WGS	100	49,922,145	99,844,279	95.95%	95.95%	99,812,358	2.8	1.18	0.84%	2.85%	18.44	-0.41	0.51	-	-
Alive	2624	25	7	2	2	2	WGS	100	25,252,683	50,505,366	93.19%	93.19%	49,921,221	2.4	1.11	0.60%	1.45%	11.04	-2.27	0.22	0.06	0.15
Deceased	2625	25	19	2	2	2	WGS	100	45,413,117	90,826,234	97.01%	97.01%	90,801,290	2.7	1.18	0.86%	3.49%	18.24	-0.64	0.41	-	-
Alive	2626	25	5	2	2	2	WGS	100	36,171,027	72,342,054	93.91%	93.91%	71,538,947	3.0	1.12	0.73%	2.44%	14.71	-3.35	0.17	-	-
Deceased	2627	25	13	2	2	2	WGS	100	64,131,646	128,263,291	90.24%	90.24%	127,918,872	2.8	1.20	0.79%	2.14%	14.05	-1.16	0.65	0.43	0.89
Deceased	2628	25	12	2	2	2	WGS	100	42,822,548	85,645,096	93.26%	93.26%	71,868,955	2.2	1.18	0.79%	2.44%	18.80	-1.68	0.47	0.85	0.85
Alive	2629	25	2	2	2	2	WGS	100	41,104,110	82,208,220	94.17%	94.17%	80,112,208	2.8	1.11	0.89%	2.84%	27.24	-1.48	0.51	0.18	0.15
Deceased	2630	25	2	2	2	2	WGS	100	48,453,471	96,906,942	94.43%	94.43%	96,131,148	2.7	1.03	0.60%	2.43%	18.16	-0.33	0.80	0.06	0.15
Alive	2631	25	2	2	2	2	WGS	100	33,683,081	67,366,161	95.31%	95.31%	67,125,285	2.0	1.24	0.63%	1.82%	18.20	-0.35	0.45	0.06	0.17
Deceased	2632	25	12	2	2	2	WGS	100	31,124,527	62,249,054	93.51%	93.51%	62,104,114	2.1	1.15	0.79%	1.74%	15.61	-0.35	0.19	0.00	0.59
Deceased	2633	25	12	2	2	2	WGS	100	31,124,526	62,249,053	93.51%	93.51%	62,104,113	2.1	1.15	0.79%	1.74%	15.61	-0.35	0.19	0.00	0.59
Alive	2634	25	1	2	2	2	WGS	100	42,126,114	84,252,228	95.77%	95.77%	83,951,221	2.8	1.50	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2635	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2636	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Deceased	2637	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2638	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2639	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Deceased	2640	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2641	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2642	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Deceased	2643	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2644	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Deceased	2645	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2646	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Deceased	2647	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2648	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Deceased	2649	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2650	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Deceased	2651	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2652	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Deceased	2653	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2654	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Deceased	2655	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2656	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Deceased	2657	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2658	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Deceased	2659	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2660	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Deceased	2661	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2662	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Deceased	2663	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2664	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Deceased	2665	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2666	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Deceased	2667	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%	3.84%	17.95	-1.45	0.52	-	-
Alive	2668	25	23	2	2	2	WGS	100	58,273,133	116,546,266	94.47%	94.47%	116,546,266	2.8	1.49	0.88%						

Supplementary Table 4. Summary of high coverage cDNA samples and genomic analyses

Patient ID	Patient type	Sequencing approach	Read type	Read length	Number of sequenced fragments	Number of sequenced reads	Percent of Q20 bases	Percent of Q30 bases	Reads mapped to genome	Whole genome coverage	Estimated 8-oxo-dG level
CSP110278P	No cancer	WGS	Paired-end	100	111,957,591	223,915,182	97.87%	94.09%	167,211,701	5.4	1.21
CSP110278P	No cancer	WGS	Paired-end	100	114,037,799	228,075,596	97.09%	92.75%	166,779,614	5.4	1.29
CSP110294P	Lung-Adenocarcinoma	WGS	Paired-end	100	214,120,371	428,240,742	97.15%	92.91%	313,975,177	10.1	1.14
CSP110304P	No cancer	WGS	Paired-end	100	232,622,449	465,244,898	97.75%	93.99%	333,577,133	10.8	1.37
CSP110308P	Lung-Adenocarcinoma	WGS	Paired-end	100	259,238,568	518,477,136	96.95%	92.24%	362,085,707	11.7	1.59
CSP110315P	No cancer	WGS	Paired-end	100	180,302,209	360,604,556	95.08%	88.75%	219,144,799	7.1	1.90
CSP110332P	No cancer	WGS	Paired-end	100	482,745,711	964,491,422	96.35%	91.65%	598,985,989	19.3	1.51
CSP110336P	Lung-SCC	WGS	Paired-end	100	235,944,337	471,888,674	96.36%	92.39%	346,124,316	11.2	1.48
CSP110344P	Lung-SCC	WGS	Paired-end	100	197,441,771	394,883,542	96.34%	90.57%	257,601,476	8.3	1.22
CSP110349P	No cancer	WGS	Paired-end	100	288,592,163	577,184,326	97.39%	93.06%	401,504,050	13.0	1.99
CSP110347P	Skin-Melanoma	WGS	Paired-end	100	264,061,871	528,123,742	96.59%	92.33%	360,599,585	11.6	1.18
CSP110359P	No cancer	WGS	Paired-end	100	129,620,634	259,241,268	97.52%	93.80%	196,325,407	6.3	1.08
CSP110358P	No cancer	WGS	Paired-end	100	288,934,085	577,868,070	97.02%	93.05%	417,887,119	13.5	1.87
CSP110362P	No cancer	WGS	Paired-end	100	188,851,782	377,703,524	97.07%	92.91%	246,193,720	8.0	2.03
CSP110367P	No cancer	WGS	Paired-end	100	129,530,151	259,060,302	97.17%	92.91%	190,012,958	6.1	2.25
CSP110372P	No cancer	WGS	Paired-end	100	262,644,051	525,288,102	97.38%	92.91%	381,528,987	12.3	1.78
CSP110387P	No cancer	WGS	Paired-end	100	165,537,915	331,075,832	96.55%	91.79%	228,174,881	7.4	1.27
CSP110401P	No cancer	WGS	Paired-end	100	233,104,285	466,208,570	97.00%	92.81%	335,659,502	10.8	1.58
CSP110403P	Lung-Adenocarcinoma	WGS	Paired-end	100	273,329,638	546,659,276	97.46%	93.65%	400,084,110	12.9	1.49
CSP110409P	No cancer	WGS	Paired-end	100	125,590,262	251,180,524	97.33%	92.65%	184,167,308	5.9	1.92
CSP110407P	No cancer	WGS	Paired-end	100	111,635,066	223,270,132	96.82%	91.87%	165,486,613	5.3	1.49
CSP110413P	No cancer	WGS	Paired-end	100	166,704,737	333,409,474	97.11%	92.39%	230,899,906	7.5	1.67
CSP110414P	Lung-Adenocarcinoma	WGS	Paired-end	100	271,065,301	542,130,602	96.45%	91.47%	389,389,382	12.6	2.34
CSP110424P	Lung-SCC	WGS	Paired-end	100	211,646,960	423,293,920	97.44%	93.08%	306,827,015	9.9	1.67
CSP110432P	No cancer	WGS	Paired-end	100	184,677,106	369,354,212	95.99%	90.55%	261,769,610	8.5	1.57
CSP110442P	No cancer	WGS	Paired-end	100	536,917,688	1,073,835,276	97.20%	92.77%	769,204,282	24.8	1.66
CSP110446P	No cancer	WGS	Paired-end	100	213,921,280	427,842,560	96.60%	92.16%	302,555,122	10.0	1.61
CSP110452P	Lung-Adenocarcinoma	WGS	Paired-end	100	233,859,246	467,718,492	96.48%	91.23%	296,331,349	9.6	1.79
CSP110454P	No cancer	WGS	Paired-end	100	547,702,301	1,095,404,602	97.49%	93.35%	802,169,883	25.9	1.75
CSP110458P	Lung-SCC	WGS	Paired-end	100	201,696,335	403,392,650	96.54%	92.27%	279,676,358	9.0	1.57
CSP110469P	Lung-Adenocarcinoma	WGS	Paired-end	100	235,511,074	511,022,148	95.67%	89.09%	317,945,299	10.3	1.40
CSP110474P	No cancer	WGS	Paired-end	100	129,875,218	259,750,436	96.45%	91.19%	191,316,533	6.2	1.42
CSP110489P	Lung-SCC	WGS	Paired-end	100	258,320,791	516,641,402	96.19%	90.55%	333,009,780	10.8	1.99
CSP110493P	No cancer	WGS	Paired-end	100	180,961,885	361,723,770	95.63%	90.00%	248,609,427	8.0	1.63
CSP110499P	No cancer	WGS	Paired-end	100	110,862,313	221,724,626	97.42%	93.38%	168,954,669	5.5	1.67
CSP110498P	No cancer	WGS	Paired-end	100	212,160,087	424,320,114	97.75%	93.81%	309,103,626	10.0	0.89
CSP110497P	No cancer	WGS	Paired-end	100	262,108,163	524,216,326	97.60%	93.71%	389,815,972	12.6	1.39
CSP110499P	No cancer	WGS	Paired-end	100	403,179,944	806,347,968	97.32%	93.06%	580,407,122	18.7	1.58

Supplementary Table 4. Summary of high coverage cDNA samples and genomic analyses

Patient ID	Patient type	Sequencing approach	Read type	Read length	Number of sequenced fragments	Number of sequenced reads	Percent of Q20 bases	Percent of Q30 bases	Reads mapped to genome	Whole genome coverage	Estimated 8-oxo-dG level
CGPLU501P	Lung-Adenocarcinoma	WGS	Paired-end	100	245,442,315	490,885,630	97.54%	93.45%	353,156,858	11.4	1.02
CGPLU504P	Lung-Adenocarcinoma	WGS	Paired-end	100	237,335,015	474,670,030	97.33%	93.06%	332,274,133	10.7	1.31
CGPLU505P	Lung-Adenocarcinoma	WGS	Paired-end	100	270,435,688	540,871,376	96.08%	89.98%	337,655,828	10.6	0.97
CGPLU509P	No cancer	WGS	Paired-end	100	225,668,024	451,736,048	97.47%	93.39%	332,857,454	10.8	2.53
CGPLU519P	No cancer	WGS	Paired-end	100	146,724,698	293,449,396	97.44%	93.21%	214,290,983	6.9	1.80
CGPLU529P	Lymphoid-B-NHL	WGS	Paired-end	100	308,682,852	617,365,704	97.31%	93.36%	445,142,405	14.4	1.43
CGPLU539P	No cancer	WGS	Paired-end	100	222,479,680	444,959,360	97.08%	92.25%	315,954,289	10.2	1.20
CGPLU541P	No cancer	WGS	Paired-end	100	157,557,222	315,114,444	97.47%	93.41%	235,286,862	7.6	1.08
CGPLU543P	Lung-Adenocarcinoma	WGS	Paired-end	100	203,760,829	407,521,658	97.17%	93.03%	309,428,859	10.0	1.21
CGPLU551P	No cancer	WGS	Paired-end	100	136,793,991	273,587,982	97.19%	92.92%	199,419,299	6.4	1.22
CGPLU557P	Skin-Melanoma	WGS	Paired-end	100	211,095,530	422,191,060	97.62%	93.67%	310,392,660	10.0	1.09
CGPLU567P	No cancer	WGS	Paired-end	100	137,537,985	275,075,970	97.03%	93.00%	199,358,378	6.4	1.03
CGPLU573P	No cancer	WGS	Paired-end	100	161,756,112	323,512,224	95.35%	89.54%	213,019,380	6.9	1.22
CGPLU581P	No cancer	WGS	Paired-end	100	140,782,605	283,565,210	97.31%	92.84%	201,297,726	6.5	1.38
CGPLU589P	No cancer	WGS	Paired-end	100	184,548,262	368,096,524	97.20%	92.75%	276,227,570	8.9	1.01
CGPLU592P	No cancer	WGS	Paired-end	100	260,993,299	521,986,598	96.13%	90.96%	368,209,814	11.9	1.31
CGPLU596P	No cancer	WGS	Paired-end	100	270,677,726	541,355,452	96.11%	91.08%	372,038,938	12.0	1.25
CGPLU599P	No cancer	WGS	Paired-end	100	133,952,771	267,905,542	97.31%	93.04%	197,941,456	6.4	1.27
CGPLU610P	No cancer	WGS	Paired-end	100	169,086,140	338,172,280	96.95%	92.64%	241,148,969	7.8	1.00
CGPLU635P	No cancer	WGS	Paired-end	100	273,761,253	547,522,506	96.50%	91.54%	402,442,860	13.0	1.25

Supplementary Table 5. Summary of ctDNA samples and genomic analyses of lung cancer patients undergoing targeted therapy*

Patient ID	Histological diagnosis	Age	Sex	Stage	Smoking status	Pack years	Treatment	Days to initial RECIST assessment	RECIST percent change in target lesions	Initial radiographic response type	Days to RECIST progression	Days on Treatment	Overall response at last follow-up	Status at last follow-up
COP11316	Mixed	50	Female	IIIA	Former smoker	15	Osimertinib	42	NA	Progressive Disease	42	73	Progressive disease	Alive
COP110209	Adenocarcinoma	52	Female	IV	Former smoker	8	Marvalertinib	25	-5.9	Stable Disease	NA	47	Progressive disease	Deceased
COP110271	Adenocarcinoma	73	Male	IV	Former smoker	NA	Marvalertinib	36	-0.2	Stable Disease	246	273	Progressive disease	Deceased
COP11088	Adenocarcinoma	59	Male	IIIA	Former smoker	NA	Osimertinib	35	-40.7	Partial Response	541	548	Progressive disease	Deceased
COP11089	Adenocarcinoma	54	Female	IV	Former smoker	NA	Osimertinib	37	-41.2	Partial Response	280	427	Progressive disease	Deceased

*Clinical, coverage, and mutant allele fraction information for patients undergoing targeted therapy were obtained from Phallen et al.³¹ and Cristiano et al.³²

Supplementary Table 5. Summary of ctDNA samples and genomic analyses of lung cancer patients undergoing targeted therapy* CONTINUED

Timepoint (days relative to treatment initiation)	Sequencing approach	Read type	Read length	Number of sequenced fragments	Number of sequenced reads	Percent of Q20 bases	Percent of Q30 bases	Reads mapped to genome	Whole genome coverage	Estimated 8- mer-level coverage	GENIVI score(=4)	Maximum mutant allele fraction from targeted sequencing
Pre-treatment, Day -53	WGS	Paired-end	100	55,769,667	111,539,794	92.65%	87.38%	103,327,297	3.3	3.40	0.53	15.72%
Pre-treatment, Day -4	WGS	Paired-end	100	42,612,304	85,224,768	92.51%	87.21%	78,087,227	2.5	3.09	0.50	45.67%
Post-treatment, Day 10	WGS	Paired-end	100	43,916,255	87,832,510	92.78%	87.59%	81,360,304	2.6	3.01	0.46	33.38%
Post-treatment, Day 37	WGS	Paired-end	100	50,967,948	101,935,896	92.57%	87.28%	94,106,646	3.0	2.58	0.50	68.01%
Pre-treatment, Day 0	WGS	Paired-end	100	50,086,599	100,173,198	89.27%	83.53%	88,595,949	2.9	3.58	0.48	0.10%
Post-treatment, Day 9	WGS	Paired-end	100	48,974,695	97,949,390	87.28%	80.96%	82,973,303	2.7	3.54	0.10	0%
Post-treatment, Day 28	WGS	Paired-end	100	45,989,749	91,979,498	89.44%	83.56%	81,787,487	2.6	3.30	0.45	0%
Pre-treatment, Day 0	WGS	Paired-end	100	68,496,676	136,993,352	94.05%	90.52%	124,970,659	4.0	3.27	0.88	3.36%
Post-treatment, Day 6	WGS	Paired-end	100	49,313,763	98,627,526	89.14%	83.16%	87,587,289	2.8	3.47	0.77	0.17%
Post-treatment, Day 20	WGS	Paired-end	100	47,121,118	94,242,236	88.82%	82.90%	81,619,670	2.6	1.53	0.66	0.18%
Post-treatment, Day 104	WGS	Paired-end	100	49,089,804	98,179,608	88.54%	81.90%	83,580,303	2.7	1.50	0.83	0.33%
Post-treatment, Day 259	WGS	Paired-end	100	44,030,080	88,060,160	90.39%	85.34%	79,973,020	2.6	1.52	0.75	0.44%
Pre-treatment, Day 0	WGS	Paired-end	100	43,827,775	87,655,550	93.01%	87.93%	81,537,455	2.6	1.33	0.82	12.17%
Post-treatment, Day 7	WGS	Paired-end	100	44,838,615	89,677,230	93.05%	88.07%	82,955,060	2.7	1.22	0.42	0.24%
Post-treatment, Day 287	WGS	Paired-end	100	58,655,454	102,310,908	92.17%	86.95%	92,287,948	3.0	1.36	0.87	2.42%
Pre-treatment, Day 0	WGS	Paired-end	100	43,230,041	86,460,082	91.82%	86.03%	79,669,676	2.6	1.44	0.57	0.42%
Post-treatment, Day 7	WGS	Paired-end	100	40,433,167	80,866,334	91.36%	85.57%	73,240,888	2.4	1.35	0.38	0.28%
Post-treatment, Day 22	WGS	Paired-end	100	34,791,256	69,582,512	93.28%	88.58%	63,652,033	2.1	1.28	0.85	0.26%

Supplementary Table 6. Summary of validation cohort and genomic analyses														
Patient ID	Collection	Primary site	Stage	Sex	Age at collection	Years from smoking	Smoking approach	Read length	Number of sequenced fragments	Number of sequenced reads	Percent of 50x reads	Percent of 100x reads	Whole genome coverage	Estimated read depth
C08210603	REC2019	Benign	NA	Male	55	0	WGS	100	38,341,425	66,822,880	93.87%	93.70%	61,513,216	1.0
C08210619	REC2019	Lung cancer	NA	Male	65	0	WGS	100	36,882,000	73,370,992	93.47%	93.20%	60,232,592	1.1
C08210622	REC2019	Lung cancer	NA	Male	70	0	WGS	100	33,630,688	67,260,272	95.70%	92.58%	62,460,736	1.0
C08210639	REC2019	Lung cancer	NA	Female	64	NA	WGS	100	29,287,273	58,574,766	96.05%	94.08%	56,827,584	1.1
C08210649	REC2019	Lung cancer	NA	Male	66	NA	WGS	100	35,196,341	70,392,682	95.70%	93.08%	69,924,592	1.1
C08210653	REC2019	Lung cancer	NA	Male	72	NA	WGS	100	31,797,250	63,594,500	96.25%	93.00%	61,020,624	1.0
C08210663	REC2019	Lung cancer	NA	Male	70	NA	WGS	100	36,491,735	72,983,470	95.97%	93.47%	68,557,966	1.1
C08210678	REC2019	Lung cancer	NA	Male	69	NA	WGS	100	35,784,696	71,569,392	96.90%	94.90%	70,830,592	1.1
C08210689	REC2019	Benign	NA	Male	74	NA	WGS	100	34,081,835	68,163,666	95.00%	93.70%	66,347,568	1.1
C08210699	REC2019	Lung cancer	NA	Male	75	NA	WGS	100	29,215,035	58,430,070	96.15%	93.00%	56,864,288	1.1
C08210709	REC2019	Lung cancer	NA	Female	74	NA	WGS	100	35,073,036	70,146,072	96.00%	94.70%	69,894,384	1.1
C08210719	REC2019	Benign	NA	Female	70	NA	WGS	100	31,013,281	62,026,562	94.00%	91.70%	58,766,864	1.1
C08210733	REC2019	Benign	NA	Male	86	NA	WGS	100	32,759,006	65,518,012	94.00%	91.60%	61,130,256	1.1
C08210739	REC2019	Benign	NA	Male	84	NA	WGS	100	32,471,262	64,942,524	96.70%	94.40%	63,594,036	1.1
C08210759	REC2019	Benign	NA	Male	65	NA	WGS	100	35,351,336	70,702,672	96.41%	93.83%	71,137,008	1.1
C08210769	REC2019	Lung cancer	NA	Male	72	NA	WGS	100	29,408,045	58,816,086	96.11%	94.08%	57,632,192	1.1
C08210779	REC2019	Benign	NA	Male	60	NA	WGS	100	33,193,847	66,387,694	96.60%	93.95%	61,940,480	1.1
C08210783	REC2019	Lung cancer	NA	Male	62	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08210799	REC2019	Benign	NA	Male	70	NA	WGS	100	32,481,994	64,963,988	96.74%	94.07%	62,940,944	1.1
C08210809	REC2019	Lung cancer	NA	Female	66	NA	WGS	100	32,421,004	64,842,008	96.75%	94.48%	61,887,488	1.1
C08210819	REC2019	Lung cancer	NA	Female	79	NA	WGS	100	31,594,457	63,188,914	95.32%	92.89%	59,286,352	1.1
C08210829	REC2019	Benign	NA	Male	57	NA	WGS	100	33,190,354	66,380,706	96.60%	93.57%	66,446,512	1.1
C08210839	REC2019	Lung cancer	NA	Male	60	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08210849	REC2019	Lung cancer	NA	Male	62	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08210859	REC2019	Lung cancer	NA	Male	64	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08210869	REC2019	Lung cancer	NA	Male	66	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08210879	REC2019	Lung cancer	NA	Male	68	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08210889	REC2019	Lung cancer	NA	Male	70	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08210899	REC2019	Lung cancer	NA	Male	72	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08210909	REC2019	Lung cancer	NA	Male	74	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08210919	REC2019	Lung cancer	NA	Male	76	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08210929	REC2019	Lung cancer	NA	Male	78	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08210939	REC2019	Lung cancer	NA	Male	80	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08210949	REC2019	Lung cancer	NA	Male	82	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08210959	REC2019	Lung cancer	NA	Male	84	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08210969	REC2019	Lung cancer	NA	Male	86	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08210979	REC2019	Lung cancer	NA	Male	88	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08210989	REC2019	Lung cancer	NA	Male	90	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08210999	REC2019	Lung cancer	NA	Male	92	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211009	REC2019	Lung cancer	NA	Male	94	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211019	REC2019	Lung cancer	NA	Male	96	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211029	REC2019	Lung cancer	NA	Male	98	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211039	REC2019	Lung cancer	NA	Male	100	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211049	REC2019	Lung cancer	NA	Male	102	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211059	REC2019	Lung cancer	NA	Male	104	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211069	REC2019	Lung cancer	NA	Male	106	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211079	REC2019	Lung cancer	NA	Male	108	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211089	REC2019	Lung cancer	NA	Male	110	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211099	REC2019	Lung cancer	NA	Male	112	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211109	REC2019	Lung cancer	NA	Male	114	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211119	REC2019	Lung cancer	NA	Male	116	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211129	REC2019	Lung cancer	NA	Male	118	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211139	REC2019	Lung cancer	NA	Male	120	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211149	REC2019	Lung cancer	NA	Male	122	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211159	REC2019	Lung cancer	NA	Male	124	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211169	REC2019	Lung cancer	NA	Male	126	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211179	REC2019	Lung cancer	NA	Male	128	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211189	REC2019	Lung cancer	NA	Male	130	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211199	REC2019	Lung cancer	NA	Male	132	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211209	REC2019	Lung cancer	NA	Male	134	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211219	REC2019	Lung cancer	NA	Male	136	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211229	REC2019	Lung cancer	NA	Male	138	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211239	REC2019	Lung cancer	NA	Male	140	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211249	REC2019	Lung cancer	NA	Male	142	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211259	REC2019	Lung cancer	NA	Male	144	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211269	REC2019	Lung cancer	NA	Male	146	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211279	REC2019	Lung cancer	NA	Male	148	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211289	REC2019	Lung cancer	NA	Male	150	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211299	REC2019	Lung cancer	NA	Male	152	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211309	REC2019	Lung cancer	NA	Male	154	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211319	REC2019	Lung cancer	NA	Male	156	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211329	REC2019	Lung cancer	NA	Male	158	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211339	REC2019	Lung cancer	NA	Male	160	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211349	REC2019	Lung cancer	NA	Male	162	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211359	REC2019	Lung cancer	NA	Male	164	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211369	REC2019	Lung cancer	NA	Male	166	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211379	REC2019	Lung cancer	NA	Male	168	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211389	REC2019	Lung cancer	NA	Male	170	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211399	REC2019	Lung cancer	NA	Male	172	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211409	REC2019	Lung cancer	NA	Male	174	NA	WGS	100	35,945,004	71,890,008	96.92%	94.10%	67,902,224	1.1
C08211419														

[illegible]

Patient ID	Collection Center / Center / Coordinator	Patient type	Histological diagnosis	Age at collection	Sex	Stage	Smoking status	Risk factor	Years from pathologic approach	Year	Sequencing approach	Read length	Number of required sequencing reads	Number of required reads	Percent of 150 bases	Percent bases	Bases mapped to genome	Whole genome coverage	Estimated read	Sample	GBM1 / GBM2				GBM3 / GBM4			
																					score	score	score	score	score	score	score	score
GBM001001	GBM	Low-grade	astrocytoma	61	Male	II	Current smoker	21	0	1982	Paired-end	100	15,151,112	17,142,144	95.85%	95.85%	93,085,490	101	0.1	100	100	100	100	100	100	100	100	
GBM001002	GBM	Low-grade	adenocarcinoma	70	Female	II	Current smoker	14.5	0	1983	Paired-end	100	14,629,031	16,257,682	96.23%	96.23%	94,613,169	1.6	1.6	100	100	100	100	100	100	100	100	
GBM001003	GBM	High-grade	glioma	69	Female	III	Former smoker	NA	14	1983	Paired-end	100	15,709,129	16,248,253	96.20%	96.20%	94,704,465	1.1	1.1	100	100	100	100	100	100	100	100	
GBM001004	GBM	High-grade	glioma	73	Male	III	Current smoker	NA	0	1982	Paired-end	100	15,342,168	17,164,330	96.13%	96.13%	93,331,595	2.2	2.2	100	100	100	100	100	100	100	100	
GBM001005	GBM	Low-grade	adenocarcinoma	67	Female	II	Former smoker	NA	10	1982	Paired-end	100	15,151,947	16,463,884	95.23%	95.23%	94,163,783	1.8	1.8	100	100	100	100	100	100	100	100	
GBM001006	GBM	Low-grade	adenocarcinoma	66	Female	II	Former smoker	NA	5	1982	Paired-end	100	15,151,843	16,161,866	96.43%	96.43%	93,018,256	0.7	0.7	100	100	100	100	100	100	100	100	
GBM001007	GBM	Low-grade	adenocarcinoma	74	Female	II	Current smoker	NA	0	1983	Paired-end	100	15,168,882	16,216,238	96.13%	96.13%	94,444,329	1.4	1.4	100	100	100	100	100	100	100	100	
GBM001008	GBM	High-grade	glioma	64	Female	III	Former smoker	NA	NA	1983	Paired-end	100	17,003,844	17,774,808	96.21%	96.21%	102,768,292	1.8	1.8	100	100	100	100	100	100	100	100	
GBM001009	GBM	Low-grade	adenocarcinoma	60	Male	II	Former smoker	1	30	1982	Paired-end	100	15,145,968	16,103,930	96.20%	96.20%	93,677,485	1.1	1.1	100	100	100	100	100	100	100	100	
GBM001010	GBM	Low-grade	adenocarcinoma	69	Female	II	Current smoker	45	0	1982	Paired-end	100	15,485,476	16,916,555	95.96%	95.96%	101,344,688	0.6	0.6	100	100	100	100	100	100	100	100	

Supplementary Table 7. Summary of SCLC vs. NSCLC analyses

Patient ID	Cohort	Lung cancer subtype	Regional difference in single molecule C>A frequency (per Mb of evaluable bases)
CGPLU283P	LUCAS	NSCLC	-4.50
CGPLU284P	LUCAS	NSCLC	9.62
CGPLU290P	LUCAS	NSCLC	-0.89
CGPLU294P	LUCAS	NSCLC	-2.89
CGPLU302P	LUCAS	NSCLC	-2.16
CGPLU306P	LUCAS	NSCLC	-2.50
CGPLU307P	LUCAS	NSCLC	-3.67
CGPLU314P	LUCAS	NSCLC	-3.31
CGPLU316P	LUCAS	NSCLC	-0.41
CGPLU320P	LUCAS	NSCLC	-5.60
CGPLU321P	LUCAS	NSCLC	0.22
CGPLU323P	LUCAS	NSCLC	0.42
CGPLU325P	LUCAS	NSCLC	-0.79
CGPLU327P	LUCAS	SCLC	8.49
CGPLU331P	LUCAS	NSCLC	6.38
CGPLU333P	LUCAS	NSCLC	-2.98
CGPLU335P	LUCAS	NSCLC	4.47
CGPLU336P	LUCAS	NSCLC	-2.72
CGPLU341P	LUCAS	NSCLC	-3.15
CGPLU343P	LUCAS	SCLC	14.26
CGPLU344P	LUCAS	NSCLC	2.14
CGPLU349P	LUCAS	NSCLC	-3.19
CGPLU369P	LUCAS	NSCLC	-3.62
CGPLU375P	LUCAS	NSCLC	-2.02
CGPLU376P	LUCAS	NSCLC	-2.30
CGPLU378P	LUCAS	NSCLC	-3.77
CGPLU384P	LUCAS	NSCLC	-3.54
CGPLU388P	LUCAS	NSCLC	-0.55
CGPLU387P	LUCAS	SCLC	-2.79
CGPLU398P	LUCAS	NSCLC	0.34
CGPLU403P	LUCAS	NSCLC	-2.20
CGPLU404P	LUCAS	SCLC	6.94

Supplementary Table 7. Summary of SCLC vs. NSCLC analyses

Patient ID	Cohort	Lung cancer subtype	Regional difference in single molecule C>A frequency (per Mb of evaluable bases)
CGPLU410P	LUCAS	NSCLC	-3.02
CGPLU412P	LUCAS	NSCLC	0.74
CGPLU414P	LUCAS	NSCLC	1.52
CGPLU424P	LUCAS	NSCLC	0.16
CGPLU426P	LUCAS	SCLC	0.42
CGPLU435P	LUCAS	NSCLC	-2.99
CGPLU441P	LUCAS	NSCLC	-3.10
CGPLU443P	LUCAS	NSCLC	-3.06
CGPLU449P	LUCAS	NSCLC	-1.90
CGPLU452P	LUCAS	NSCLC	-1.17
CGPLU456P	LUCAS	NSCLC	-2.05
CGPLU457P	LUCAS	SCLC	5.73
CGPLU458P	LUCAS	SCLC	12.65
CGPLU467P	LUCAS	NSCLC	-2.39
CGPLU468P	LUCAS	NSCLC	-5.75
CGPLU469P	LUCAS	NSCLC	-6.37
CGPLU473P	LUCAS	SCLC	0.15
CGPLU477P	LUCAS	NSCLC	-5.68
CGPLU480P	LUCAS	NSCLC	-1.73
CGPLU483P	LUCAS	NSCLC	-1.26
CGPLU488P	LUCAS	NSCLC	-0.86
CGPLU501P	LUCAS	NSCLC	-0.78
CGPLU504P	LUCAS	NSCLC	-0.20
CGPLU505P	LUCAS	NSCLC	10.21
CGPLU506P	LUCAS	NSCLC	-4.73
CGPLU518P	LUCAS	NSCLC	1.27
CGPLU526P	LUCAS	NSCLC	10.07
CGPLU531P	LUCAS	SCLC	3.50
CGPLU533P	LUCAS	NSCLC	-2.47
CGPLU538P	LUCAS	NSCLC	-1.87
CGPLU543P	LUCAS	NSCLC	0.29
CGPLU544P	LUCAS	NSCLC	-1.01

Supplementary Table 7. Summary of SCLC vs. NSCLC analyses

Patient ID	Cohort	Lung cancer subtype	Regional difference in single molecule C>A frequency (per Mb of evaluable bases)
CGPLLU550P	LUCAS	NSCLC	-0.41
CGPLLU558P	LUCAS	NSCLC	-0.12
CGPLLU559P	LUCAS	NSCLC	-6.35
CGPLLU565P	LUCAS	NSCLC	-4.28
CGPLLU566P	LUCAS	NSCLC	-2.47
CGPLLU569P	LUCAS	NSCLC	-3.19
CGPLLU571P	LUCAS	NSCLC	-0.40
CGPLLU578P	LUCAS	NSCLC	-3.22
CGPLLU585P	LUCAS	NSCLC	-1.88
CGPLLU587P	LUCAS	NSCLC	-0.80
CGPLLU584P	LUCAS	SCLC	4.72
CGPLLU586P	LUCAS	NSCLC	-3.11
CGPLLU607P	LUCAS	NSCLC	-2.94
CGPLLU611P	LUCAS	NSCLC	-3.60
CGPLLU616P	LUCAS	NSCLC	-2.34
CGPLLU622P	LUCAS	NSCLC	-6.88
CGPLLU626P	LUCAS	NSCLC	-7.10
CGPLLU627P	LUCAS	NSCLC	-3.94
CGPLLU633P	LUCAS	NSCLC	-5.13
CGPLLU636P	LUCAS	NSCLC	2.36
CGPLLU637P	LUCAS	NSCLC	-2.61
CGPLLU641P	Validation	SCLC	-2.10
CGPLLU642P	Validation	NSCLC	-1.18
CGPLLU644P	Validation	NSCLC	0.62
CGPLLU645P	Validation	NSCLC	-4.36
CGPLLU646P	Validation	NSCLC	-5.50
CGPLLU649P	Validation	SCLC	-0.97
CGPLLU650P	Validation	NSCLC	-1.73
CGPLLU658P	Validation	NSCLC	-1.29
CGPLLU661P	Validation	SCLC	7.42
CGPLLU670P	Validation	NSCLC	-3.97
CGPLLU672P_2	Validation	NSCLC	-3.77

Supplementary Table 7. Summary of SCLC vs. NSCLC analyses

Patient ID	Cohort	Lung cancer subtype	Regional difference in single molecule C>A frequency (per Mb of evaluable bases)
CGPLLU675P	Validation	NSCLC	-5.83
CGPLLU683P	Validation	NSCLC	-3.45
CGPLLU684P	Validation	NSCLC	-2.16
CGPLLU686P	Validation	NSCLC	-4.93
CGPLLU688P	Validation	NSCLC	-3.52
CGPLLU690P	Validation	NSCLC	-3.21
CGPLLU693P	Validation	NSCLC	-2.07
CGPLLU696P	Validation	NSCLC	-3.88
CGPLLU697P	Validation	NSCLC	1.72
CGPLLU698P	Validation	NSCLC	-3.48
CGPLLU703P	Validation	NSCLC	-0.69
CGPLLU705P	Validation	NSCLC	-1.68
CGPLLU710P	Validation	NSCLC	-0.58
CGPLLU712P	Validation	NSCLC	0.50
CGPLLU713P	Validation	NSCLC	-2.98
CGPLLU718P	Validation	NSCLC	-3.71

Supplementary Table 6. Summary of liver cohort and genomic analyses

Patient ID	Patient type	Age	Sex	BCLC stage	Sequencing approach	Read type	Read length	Number of sequenced fragments	Number of sequenced reads	Percent of D20 bases	Percent of D30 bases	Reads mapped to genome	Whole genome coverage	Estimated B-cell level	Inter-CNA fraction	GERM score
CG-100P	HCC	70	Male	B	WG5	Paired-end	100	49,955,549	95,511,038	97.59%	94.42%	77,394,360	2.5	8.30	93.44%	0.95
CG-115P	HCC	62	Male	A	WG5	Paired-end	100	42,991,733	85,383,158	97.95%	94.27%	69,670,156	2.1	4.71	1.13%	0.97
CG-115P	HCC	50	Female	A	WG5	Paired-end	100	71,254,158	142,518,336	97.89%	94.53%	107,371,330	3.5	5.94	1.57%	0.78
CG-115P	HCC	50	Male	C	WG5	Paired-end	100	51,087,050	102,174,101	98.02%	94.50%	77,075,659	2.5	5.39	11.15%	0.78
CG-148P	HCC	64	Male	B	WG5	Paired-end	100	38,489,598	96,577,392	98.84%	92.88%	48,022,670	1.6	2.38	1.18%	0.98
CG-151P	HCC	75	Male	C	WG5	Paired-end	100	54,581,778	109,163,556	95.24%	90.88%	77,206,613	2.5	1.62	21.11%	1.00
CG-118P	HCC	55	Male	B	WG5	Paired-end	100	14,540,154	49,581,538	97.80%	95.91%	57,552,578	1.3	1.55	0.54%	0.95
CG-117P	HCC	71	Male	A	WG5	Paired-end	100	42,580,159	85,177,319	98.08%	94.58%	63,476,654	2.1	1.50	1.20%	0.71
CG-133P	HCC	69	Male	B	WG5	Paired-end	100	44,346,174	128,481,148	92.89%	88.15%	76,465,607	2.5	1.49	1.20%	0.67
CG-133P	HCC	62	Male	B	WG5	Paired-end	100	31,350,545	100,701,038	97.40%	93.55%	75,888,335	2.5	1.43	0.30%	0.95
CG-119P	HCC	57	Female	A	WG5	Paired-end	100	47,268,041	134,536,002	97.75%	94.00%	102,410,984	3.5	0.11	0.21%	0.96
CG-120P	HCC	67	Male	C	WG5	Paired-end	100	40,520,900	121,641,956	95.05%	91.75%	67,140,801	2.8	1.33	93.17%	1.00
CG-121P	HCC	68	Male	B	WG5	Paired-end	100	51,706,040	127,378,086	95.33%	90.64%	90,501,717	2.6	1.72	0.50%	0.52
CG-122P	HCC	72	Male	B	WG5	Paired-end	100	60,967,546	121,975,080	92.73%	94.11%	81,830,159	3.0	1.16	0.76%	0.56
CG-123P	HCC	67	Female	B	WG5	Paired-end	100	50,101,541	100,255,082	96.81%	92.57%	74,896,776	2.4	1.39	58.55%	0.99
CG-134P	Cirrhosis	67	Female	N	WG5	Paired-end	100	52,277,037	104,544,074	97.72%	94.00%	79,598,443	2.6	0.74	0.86%	0.42
CG-125P	Cirrhosis	57	Female	N	WG5	Paired-end	100	64,405,363	127,806,723	97.05%	94.10%	85,246,921	3.1	2.88	0.86%	0.35
CG-123P	Cirrhosis	35	Female	N	WG5	Paired-end	100	51,311,458	102,621,917	96.80%	92.75%	74,818,946	1.4	5.48	1.17%	0.50
CG-127P	Cirrhosis	40	Male	N	WG5	Paired-end	100	77,872,100	154,856,200	97.85%	95.81%	114,935,817	3.7	2.28	0.19%	0.60
CG-128P	Cirrhosis	60	Female	N	WG5	Paired-end	100	78,116,350	148,751,600	97.57%	95.57%	108,975,858	3.5	3.48	0.77%	0.62
CG-125P	Cirrhosis	57	Male	N	WG5	Paired-end	100	44,360,135	88,751,039	97.77%	94.66%	65,025,144	2.1	3.40	0.91%	0.67
CG-122P	HCC	54	Male	B	WG5	Paired-end	100	51,915,089	104,631,038	97.12%	94.89%	74,550,156	1.4	3.85	9.22%	0.77
CG-130P	Cirrhosis	50	Male	B	WG5	Paired-end	100	42,570,593	100,730,856	97.05%	94.41%	92,231,451	1.0	2.57	0.47%	0.44
CG-131P	Cirrhosis	60	Male	N	WG5	Paired-end	100	46,066,514	90,133,038	96.41%	92.21%	65,077,016	1.0	3.11	0.25%	0.15
CG-132P	Cirrhosis	51	Female	N	WG5	Paired-end	100	52,971,106	105,942,112	96.94%	92.04%	95,506,151	1.2	2.85	0.30%	0.35
CG-120P	HCC	64	Male	C	WG5	Paired-end	100	46,050,349	96,796,438	97.11%	93.92%	69,970,452	2.3	1.67	1.24%	0.62
CG-136P	Cirrhosis	62	Male	N	WG5	Paired-end	100	56,277,111	112,554,422	97.53%	93.13%	89,295,408	2.9	1.42	0.52%	0.24
CG-135P	Cirrhosis	38	Male	N	WG5	Paired-end	100	47,111,451	95,225,907	97.53%	92.57%	71,231,173	2.3	1.58	0.83%	0.67
CG-135P	Cirrhosis	55	Male	N	WG5	Paired-end	100	48,854,559	97,708,708	97.54%	93.82%	68,746,187	2.2	1.49	0.71%	0.57
CG-132P	Cirrhosis	68	Female	N	WG5	Paired-end	100	56,280,655	112,531,180	97.03%	92.97%	82,896,688	2.7	1.16	0.40%	0.63
CG-133P	Cirrhosis	65	Male	N	WG5	Paired-end	100	55,351,113	110,707,076	95.50%	93.76%	81,262,253	2.6	0.92	0.36%	0.50
CG-135P	HCC	70	Male	A	WG5	Paired-end	100	49,378,700	118,851,104	97.03%	92.06%	100,054,167	3.1	6.75	0.34%	0.55
CG-135P	HCC	61	Female	A	WG5	Paired-end	100	48,164,038	96,328,076	97.94%	94.46%	71,887,306	2.8	3.52	1.17%	0.55
CG-140P	HCC	65	Male	B	WG5	Paired-end	100	43,785,051	87,571,132	97.56%	93.54%	70,531,014	2.3	4.27	0.59%	0.66
CG-141P	HCC	54	Male	C	WG5	Paired-end	100	51,451,154	110,657,395	97.41%	93.47%	101,175,038	3.3	2.69	48.11%	0.95
CG-142P	Cirrhosis	56	Male	C	WG5	Paired-end	100	54,946,175	109,891,350	97.17%	93.12%	94,585,156	2.7	2.87	54.57%	1.00
CG-145P	HCC	45	Female	C	WG5	Paired-end	100	54,352,800	108,685,600	97.83%	93.54%	88,548,516	2.8	2.89	92.86%	0.99
CG-146P	HCC	75	Male	B	WG5	Paired-end	100	56,961,771	110,173,142	96.34%	90.81%	81,174,803	2.7	5.61	1.24%	0.88
CG-146P	HCC	86	Male	C	WG5	Paired-end	100	43,987,174	88,184,148	97.51%	93.47%	69,075,084	2.2	3.84	29.66%	0.82
CG-148P	HCC	38	Male	C	WG5	Paired-end	100	50,823,115	101,634,430	97.14%	93.05%	87,175,138	2.6	5.10	19.15%	0.83
CG-147P	HCC	70	Male	C	WG5	Paired-end	100	64,026,463	128,051,926	97.78%	94.69%	105,624,098	3.4	2.62	8.71%	0.98
CG-148P	HCC	67	Male	B	WG5	Paired-end	100	61,123,152	122,247,704	96.64%	92.51%	92,357,451	3.0	2.93	0.71%	0.87
CG-148P	HCC	62	Male	B	WG5	Paired-end	100	52,881,026	105,755,257	97.15%	93.15%	82,894,658	2.6	2.50	4.70%	0.88
CG-149P	HCC	67	Male	B	WG5	Paired-end	100	71,628,854	143,257,708	97.51%	94.36%	107,007,937	3.8	3.59	0.55%	0.68
CG-150P	HCC	70	Male	B	WG5	Paired-end	100	49,408,701	98,817,402	97.42%	93.49%	69,007,719	2.1	2.49	1.22%	0.98
CG-151P	HCC	60	Female	C	WG5	Paired-end	100	51,151,134	102,302,268	97.45%	93.95%	94,022,602	2.9	2.44	57.12%	0.97
CG-152P	HCC	65	Male	B	WG5	Paired-end	100	46,650,059	93,300,118	96.56%	92.71%	69,648,846	2.1	2.23	21.88%	1.00
CG-152P	HCC	60	Male	B	WG5	Paired-end	100	47,168,551	94,337,104	97.58%	93.75%	71,105,385	2.3	3.21	0.24%	0.95

Supplementary Table 6. Summary of liver cohort and genomic analyses

Patient ID	Patient type	Age	Sex	BCLC stage	Sequencing approach	Read type	Read length	Number of sequenced fragments	Number of sequenced reads	Percent of D20 bases	Percent of D30 bases	Reads mapped to genome	Whole genome coverage	Estimated B-cell level	Inter-CNA fraction	GERM score
CG-154P	HCC	71	Male	A	WG5	Paired-end	100	54,545,153	109,090,310	97.18%	93.15%	81,053,889	2.6	1.68	0.31%	0.52
CG-155P	HCC	53	Female	A	WG5	Paired-end	100	7,455,513	15,515,036	97.15%	93.15%	11,471,150	0.4	1.98	1.38%	1.00
CG-156P	HCC	65	Male	B	WG5	Paired-end	100	52,740,193	105,514,866	98.40%	92.05%	84,514,885	1.1	4.57	0.81%	0.76
CG-157P	HCC	65	Male	A	WG5	Paired-end	100	56,085,121	110,175,042	97.09%	93.17%	98,484,198	2.7	2.70	0.35%	0.72
CG-158P	HCC	64	Male	B	WG5	Paired-end	100	39,005,943	78,111,886	95.09%	90.72%	55,185,831	1.0	2.76	5.50%	0.80
CG-159P	HCC	64	Male	B	WG5	Paired-end	100	108,617,623	217,555,245	97.15%	93.11%	161,058,519	5.2	2.62	1.48%	0.65
CG-160P	HCC	68	Male	B	WG5	Paired-end	100	51,840,169	98,684,338	96.67%	90.96%	59,486,170	1.0	1.91	8.52%	0.69
CG-161P	HCC	67	Male	B	WG5	Paired-end	100	48,008,199	96,016,312	97.21%	93.30%	69,458,982	1.7	1.28	14.84%	0.95
CG-162P	HCC	64	Male	B	WG5	Paired-end	100	61,980,141	123,960,282	96.94%	92.09%	92,937,082	3.0	1.73	0.38%	0.69
CG-163P	HCC	64	Male	B	WG5	Paired-end	100	48,555,110	97,070,470	97.73%	94.11%	70,774,784	1.3	1.35	0.50%	0.87
CG-164P	HCC	71	Male	C	WG5	Paired-end	100	51,249,803	102,500,608	98.80%	91.85%	71,872,456	2.8	3.20	0.14%	1.00
CG-165P	HCC	55	Male	A	WG5	Paired-end	100	52,906,191	105,812,384	97.02%	94.24%	81,507,875	2.6	2.79	1.56%	0.67
CG-166P	HCC	60	Male	C	WG5	Paired-end	100	49,010,537	98,020,874	97.49%	93.94%	64,780,580	2.3	1.24	4.20%	0.98
CG-169P	HCC	70	Male	A	WG5	Paired-end	100	59,769,039	119,537,413	97.04%	94.04%	99,408,325	2.9	5.04	1.72%	0.74

CLAIMS

What is claimed:

1. A method of determining the frequency of somatic mutations in a subject, comprising:
extracting cell-free DNA (cfDNA) from a subject's biological sample;
generating genomic libraries from the extracted cfDNA;
sequencing individual cfDNA molecules to obtain mutation profiles;
determining multiregional differences in mutation profiles; and,
determining the frequency of somatic mutations in the subject.
2. The method of claim 1, wherein the determining of genome-wide mutation and fragmentation profiles comprises identifying mutations in sequences of individual cfDNA molecules and changes in fragment lengths.
3. The method of claim 1 or 2, wherein the mutation profiles comprise mutation frequency and type of mutation across the subject's genome.
4. The method of claim 3, wherein the mutation profiles across the subject's genome are determined using non-overlapping bins ranging in size from at least about one thousand bases to at least about twenty million bases.
5. The method of claim 3, wherein the mutation profiles across the subject's genome are determined using non-overlapping bins ranging in size from at least about one thousand bases to at least about ten million bases.
6. The method of claim 3, wherein the mutation profiles across the subject's genome is determined using non-overlapping bins ranging in size from at least about one thousand bases to at least about five million bases.
7. The method of claim 3, wherein mutations for each sequenced molecule are determined after removing common germline variants, and unevaluable regions.
8. The method of any one of claims 1-7, wherein the frequency of single molecule somatic mutations and type of mutation across the subject's genome is diagnostic of cancer as compared

to the frequency of single molecule somatic mutations and type of mutation across a normal subject's genome.

9. The method of any of claims 1-8 where such analysis is performed in a subject from whom tumor tissue is unavailable.
10. A method of treating cancer in a subject, the method comprising: extracting cell-free DNA (cfDNA) from a subject's biological sample;
generating genomic libraries from the extracted cfDNA;
sequencing individual cfDNA molecules to obtain mutation profiles;
determining multiregional differences in mutation profiles and determining the frequency of somatic mutations in the subject; and on the basis thereof
administering a cancer treatment to the subject.
11. The method of claim 10, wherein the cancer treatment comprises: surgery, adjuvant chemotherapy, neoadjuvant chemotherapy, radiation therapy, hormone therapy, cytotoxic therapy, immunotherapy, adoptive T cell therapy, targeted therapy, and combinations thereof.
12. The method of claim 10 or 11, wherein the cancer comprises colorectal cancer, lung cancer, breast cancer, gastric cancers, pancreatic cancers, bile duct cancers, brain cancer or ovarian cancer.
13. The method of claim 12, wherein the lung cancer is small cell lung cancer (SCLC).
14. The method of claim 12, wherein the lung cancer is non-small cell lung cancer (NSCLC).
15. The method of any one of claims 10-14, wherein subjects with cancer comprise altered mutational profiles associated with chromatin organization as compared to healthy individuals.
16. The method of any one of claims 10-15, wherein genome-wide mutation and fragmentation profiles comprises identifying mutations in sequences of individual cfDNA molecules and changes in fragment lengths.
17. The method of claim 10, wherein the mutation profiles comprise mutation frequency and type of mutation across the subject's genome.

18. The method of claim 17, wherein the mutation profiles across the subject's genome are determined using non-overlapping bins ranging in size from at least about one thousand bases to at least about twenty million bases.
19. The method of claim 17, wherein the mutation profiles across the subject's genome are determined using non-overlapping bins ranging in size from at least about one thousand bases to at least about ten million bases.
20. The method of claim 17, wherein the mutation profiles across the subject's genome determined using non-overlapping bins ranging in size from at least about one thousand bases to at least about five million bases.
21. The method of claim 16, wherein genome-wide mutations for each sequenced molecule are determined after removing common germline variants, and unevaluable regions.
22. A method of determining regional frequency of mutations across a genome comprising:
sequencing individual cfDNA molecules isolated from a subject,
estimating mutation frequencies and types of mutations across the genome;
determining the mutation types and frequencies in genomic regions altered in cancer to mutation profiles and regions mutated in normal cfDNA to determine multiregional differences in mutation profiles; thereby,
determining regional frequency of mutations across a genome.
23. The method of claim 22, wherein the estimation of mutation frequencies and types of mutations across the genome comprise using non-overlapping bins ranging in size from thousands to millions of bases.
24. The method of claims 22 or 23, wherein tumor specific changes are quantified by one or more assays.
25. The method of claim 24, wherein the one or more assays comprise *in silico* dilution assays and/or downsampling assays.
26. The method of any one of claims 22-25, wherein each sequenced molecule is scanned for single nucleotide changes after removing common germline variants and/or unevaluable regions.

27. The method of any one of claims 22-26 wherein the genomic regions are characterized by late replication timing, low gene expression, B compartmentalization, high H3K9me3 abundance, low GC content, or a combination thereof.
28. The method of any one of claims 21-26, wherein the frequency of putative mutations is defined as the number of variants per million evaluated positions across all the DNA molecules sequenced.
29. The method of any one of claims 21-28, further comprising combining mutational profiles and genome-wide fragmentation profiles.
30. The method of any one of claims 21-29, further comprising executing a machine learning model for determining changes in genome-wide mutational profiles that classifies or excludes the subject as having or at risk of having cancer based on the genome-wide mutational profile identified for the subject.
31. A method of determining whether a subject is a responder to a treatment based on the outcome of performing a method of any one of claims 1-30 or combinations thereof.
32. The method of claim 31, wherein the treatment is selected from surgery, adjuvant chemotherapy, neoadjuvant chemotherapy, radiation therapy, hormone therapy, cytotoxic therapy, immunotherapy, adoptive T cell therapy, targeted therapy, and combinations thereof.