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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
02 May 2024 (02.05.2024)

WIPO | PCT

(10) International Publication Number
WO 2024/091545 A1

(51) International Patent Classification:
C12N 15/10 (2006.01) *C12Q 1/6869* (2018.01)
C12Q 1/6806 (2018.01) *C12Q 1/6886* (2018.01)
C12Q 1/6827 (2018.01) *G16B 20/20* (2019.01)
C12Q 1/6844 (2018.01) *G16B 30/10* (2019.01)

(21) International Application Number:
PCT/US2023/035877

(22) International Filing Date:
25 October 2023 (25.10.2023)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/380,915 25 October 2022 (25.10.2022) US

(71) Applicant: **CORNELL UNIVERSITY** [US/US]; 395 Pine Tree Road, Suite 310, Ithaca, NY 14850 (US).

(72) Inventors: **LANDAU, Dan-Avi**; 882 Union Street, Apt 3B, Brooklyn, NY 11215 (US). **CHENG, Alexandre, Pellan**; 461 Dean Street #22E, Brooklyn, NY 11217 (US).

(74) Agent: **HUESTIS, Erik, A.** et al.; Foley Hoag LLP, 155 Seaport Boulevard, Boston, MA 02210-2600 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV,

WO 2024/091545 A1

(54) Title: NUCLEIC ACID ERROR SUPPRESSION

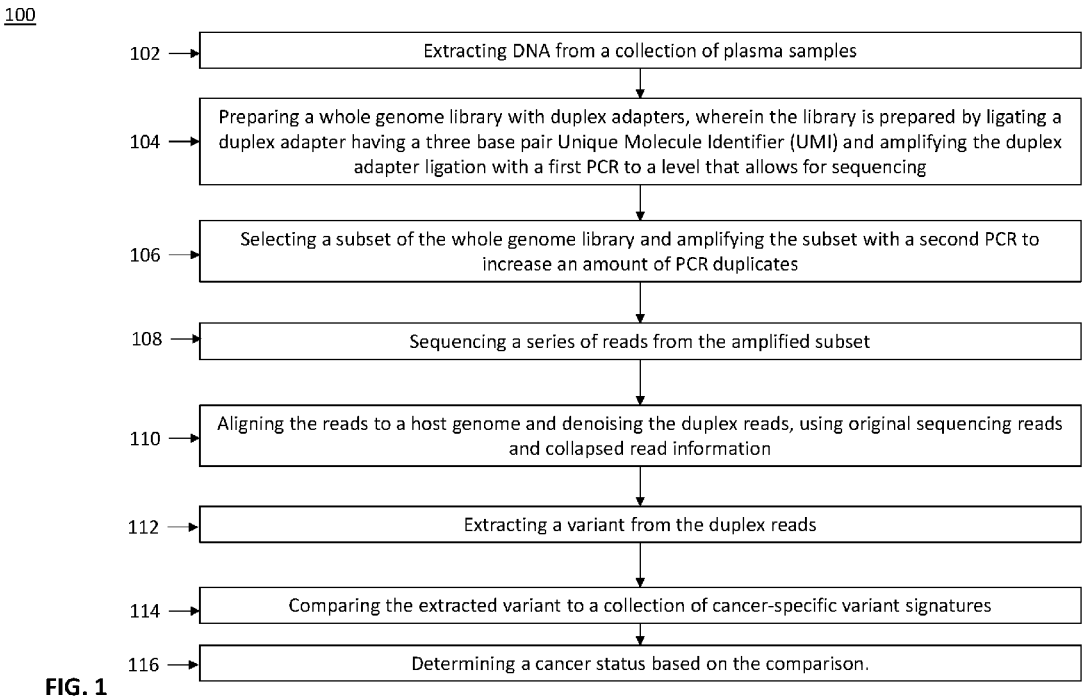


FIG. 1

(57) **Abstract:** Nucleic acid error suppression is provided. In various embodiments, DNA is extracted from a collection of plasma samples. A sequence library with duplex adapters is prepared. The library is prepared by ligating a duplex adapter having a Unique Molecule Identifier (UMI) to an end of each of a plurality of strands of the extracted DNA ard amplifying the extracted DNA with a first polymerase chain reaction (PCR). A subset of the whole genome library is selected and the subset is amplified with a second PCR to increase an amount of PCR duplicates. A plurality of duplex reads is sequenced fron the amplified subset.

[Continued on next page]

GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *of inventorship (Rule 4.17(iv))*

Published:

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

NUCLEIC ACID ERROR SUPRESSION

RELATED APPLICATION(S)

[0001] This application claims the benefit of priority to U.S. Provisional Application No. 63/380915, filed October 25, 2022, which is hereby incorporated by reference in its entirety.

TECHNICAL FIELD

[0002] The invention relates generally to the field of medical diagnostics. In particular, embodiments of the disclosure relate to methods and systems for reducing sequencing error rates in cancer detection and other fields requiring low error sequencing.

BACKGROUND

[0003] Monitoring circulating cell-free DNA (cfDNA) has been shown to be a promising clinical tool for non-invasive cancer detection. While analysis of cancer-specific epigenetic markers, such as DNA methylation and histone modifications, has been applied to cfDNA for detection of various cancers, mutation-based approaches using direct genomic sequencing of somatic variants found in circulating tumor DNA (ctDNA) afford more specificity and clinically-actionable information. As such, ctDNA genome sequencing is preferable for clinical applications, particularly in cases where there is low burden of disease, such as early cancer screening, detection of minimal residual disease (MRD) after treatment or surgery and relapse monitoring of emergent resistant mutations for guided therapy. In these scenarios, tumor fraction is low, such that robust detection requires methods with exquisite sensitivity.

[0004] Prevailing methods of ctDNA detection use targeted sequencing protocols, which increase the number of genomes sequenced at a targeted location. However, high throughput targeted sequencing rapidly exhausts available genomes for sequencing (1,000-10,000 genome equivalents (GEs) per mL of plasma), which sets a design-based ceiling on ctDNA detection, where further increasing sequencing depth at a targeted site affords no advantage after the limited number of GEs has already been sequenced. Alternatively, to overcome these limitations, whole genome sequencing (WGS) approaches exploit breadth of coverage to supplant depth, eliminating the reliance on the detection of a single site to increase ctDNA characterization in low tumor fraction settings. For example, recent method MRDetect uses

primary tumor mutational profiles to inform genome-wide tumor single nucleotide variant (SNV) detection in cfDNA, such that the available number of GEs no longer is the limiting factor for successful ctDNA detection.

[0005] The detection challenges presented by sparsity are significant, calling for broad, accurate and deep cfDNA sequencing. Thus, whole-genome, low-error, high-coverage methods are necessary for robust ctDNA analysis. However, the costs associated with these approaches are often prohibitive, particularly for clinical application. Although genome sequencing costs have rapidly dropped since the introduction of high-throughput next generation sequencing, more recently this decrease has stagnated. As such, sequencing cost is still a significant barrier to implementation of high-depth WGS for liquid biopsies, where in clinically important applications, tumor fractions are low (~10-5) and shallow WGS is insufficient for ctDNA detection. To estimate the cost of implementing WGS for successful detection of ctDNA at such fractions, the probability of detecting a single mutation in a cfDNA sample can be modeled, given the number of GEs, the tumor fraction and sequencing depth[17]. It is estimated that a sequencing depth of over 100x is required for single mutation detection in tumor-derived DNA at this low fraction level, making the per-sample cost of WGS exceedingly high for application at scale with established technologies (~2,000USD per sample using an Illumina Novaseq S4 flow cell with v1.5 reagent costs).

[0006] Recently, a new low-cost, high-throughput sequencing method utilizing mostly natural sequencing-by-synthesis (mnSBS) has been developed by Ultima Genomics. The Ultima sequencing platform produces single-end reads at ~10 billion reads per run for 1$/GB, thus substantially lowering sequencing costs compared with current platforms. This cost efficiency holds great promise for many genomics applications, and this approach has now been applied to Genome-In-A-Bottle and 1000 Genomes reference samples and adapted for single-cell RNA-seq studies. However, mnSBS/Ultima sequencing has not been harnessed for application to clinical cfDNA samples for ctDNA sequencing. In addition, notably, the error rate profiles of this new sequencing method have not been fully characterized, nor have they been rigorously compared with competing technologies. Importantly, for potential application to clinical disease monitoring of ctDNA, it is especially crucial to have accurate error rate estimates due to the high sensitivity of low-burden ctDNA detection.

[0007] Accordingly, there is a need for methods of reducing sequencing error rates in disease detection. The technical challenges imparted by the sparsity of ctDNA in low-burden disease settings may be overcome by increasing sequencing depth, accompanied by a low error rate. Unique molecular identifier (UMI) error suppression techniques or duplex

sequencing, can increase accuracy in differentiating true somatic variant calls from errors introduced by sequencing, and may be combined with deep genomic sequencing to optimize successful detection of low-burden disease in a clinical setting.

SUMMARY

[0008] Here, to investigate the utility of deep WGS for ctDNA detection, Ultima Genomics' mnSBS sequencing platform is used to sequence circulating cell-free DNA reads from plasma samples from healthy controls, cancer patients and patient-derived xenograft mouse models. In a first proof-of-principle study, it is shows that deep WGS (~100x) with in silico error correction allows ctDNA detection within the part per million range. By leveraging the cost-effective and high-throughput nature of Ultima Genomics sequencing (109 reads per run at 1$/Gb), high coverage duplex-sequencing libraries of cell-free DNA can be produced, achieving error rates as low as 2.7x10-7. This allows accurate assessment of disease burden in post-treatment melanoma patients in the absence of any tumor information. Together, this demonstrates the utility of deep WGS in clinical samples for ctDNA detection.

[0009] According to certain aspects of the present disclosure, systems and methods are disclosed for detecting cancer with a lower sequencing error rate.

[0010] In an embodiment, a method comprises extracting DNA from a collection of plasma samples; preparing a whole genome library with duplex adapters, wherein the whole genome library is prepared by ligating a duplex adapter having a Unique Molecule Identifier (UMI) to an end of each of a plurality of strands of the extracted DNA and amplifying the extracted DNA with a first PCR; selecting a subset of the whole genome library; amplifying the subset with a second PCR to increase a number of PCR duplicates; sequencing a plurality of duplex reads from the amplified subset; aligning the plurality of duplex reads to a host genome and denoising the plurality of duplex reads based on said alignment; detecting the presence of a variant in at least one of the plurality of duplex reads; determining a signature of the variant; comparing the signature of the variant to a collection of disease-specific variant signatures; and determining a disease type based on the comparison.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] The accompanying drawings, which are incorporated into and constitute a part of this specification, illustrate various exemplary embodiments and together with the description, serve to explain the principles of the disclosed embodiments.

[0012] **FIG. 1** is a flowchart illustrating a method for detecting mutation signatures to determine a cancer status, according to techniques disclosed herein.

[0013] **FIG. 2A** is a series of graphs illustrating error rates and sequencing coverage for tumor fractions at or below 10^{-5}, according to techniques disclosed herein.

[0014] **FIG. 2B** is a flowchart of a pre-analytical process to prepare a cfDNA library, according to techniques disclosed herein.

[0015] **FIG. 2C** is a graph illustrating sequencing depths for matched Illumina and Ultima datasets, according to techniques disclosed herein.

[0016] **FIG. 2D** is a comparison of normalized read coverage of a sequenced matched cfDNA sample, according to techniques disclosed herein.

[0017] **FIG. 2E** is a comparison of copy number-based variant (CNV) and single-nucleotide variant (SNV) tumor fractions, according to techniques disclosed herein.

[0018] **FIG. 2F** is a graph of an *in silico* mixing study, according to techniques disclosed herein.

[0019] **FIG. 3A** is a series of graphs illustrating duplex whole genome sequencing (WGS) on a mouse (left) and patient (right) sample, according to techniques disclosed herein.

[0020] **FIG. 3B** is a graphical comparison of variant allele frequencies calculated using unfiltered sequencing reads, according to techniques disclosed herein.

[0021] **FIG. 3C** is a graph comparing the model allele fraction of a patient with progressive disease in duplex corrected positions and copy-number based tumor fraction estimations, according to techniques disclosed herein.

[0022] **FIG. 3D** is a graph illustrating exemplary trinucleotide frequencies from a melanoma-associated UV signature.

[0023] **FIG. 3E** is a comparison of cosine similarities with either the SBS7 or the SBS1B mutation across conditions, according to techniques disclosed herein.

[0024] **FIG. 3F** is a graph illustrating a signature score and ctDNA detection of an *in silico* mixing study of metastatic melanoma samples, according to techniques disclosed herein.

[0025] **FIG. 3G** is a graph illustrating a series of signature scores of melanoma signature 7 in plasma cfDNA samples using duplex WGS, according to techniques disclosed herein.

[0026] **FIG. 3H** is a graph illustrating estimated tumor fraction of samples with elevated signature scores, according to techniques disclosed herein.

[0027] **FIG. 4** is a series of graphs illustrating frequency of cfDNA fragment lengths in single-end Ultima sequencing datasets matched with paired-end Illumina sequencing, according to techniques disclosed herein.

[0028] **FIG. 5A** is a graph illustrating the UG specific blacklist, according to techniques disclosed herein.

[0029] **FIG. 5B** is an illustration of overlap between the UG blacklist and other low confidence regions, according to techniques disclosed herein.

[0030] **FIG. 5C** is a graph illustrating the overlap between melanoma tumor tissue SNVs and low confidence regions, according to techniques disclosed herein.

[0031] **FIG. 6A** is a heatmap of cosine similarities in cancer-free samples and high-burden ctDNA samples, according to techniques disclosed herein.

[0032] **FIG. 6B** is a boxplot of cosine similarities for three correction methods in the same cancer-free samples and high-burden ctDNA samples, according to techniques disclosed herein.

[0033] **FIG. 7A** is a graph illustrating a deconvolution of duplex-corrected mutations into representative mutational signatures.

[0034] **FIG. 7B** is a correlation plot between age at cancer diagnosis and number of clock-like mutations attributed to SBS1A and SBS1B, according to techniques disclosed herein.

[0035] **FIG. 8** is a graph of a tumor-agnostic copy-number based tumor fraction estimation in cancer-free control samples and pre-surgery melanoma plasma, according to techniques disclosed herein.

[0036] **FIG. 9A** is a graph illustrating a homopolymer size between two PCR duplicates, according to techniques disclosed herein.

[0037] **FIG. 9B** is a graph illustrating a homopolymer size between a read and an aligned reference, according to techniques disclosed herein.

[0038] **FIG. 9C** is a graph illustrating frequency of homopolymer size across the human genome, according to techniques disclosed herein.

[0039] **FIG. 9D** is a graph illustrating indel calling accuracy by PCR duplicate family sizes, according to techniques disclosed herein.

[0040] **FIG. 10** is a graph illustrating a single nucleotide variant analysis of matched Ultima and Illumina sequencing datasets, according to techniques disclosed herein.

[0041] **FIG. 11A** is a flow chart of a sequencing process providing predictable, error-robust motifs, according to techniques disclosed herein.

[0042] **FIG. 11B** is a graph of error rate by sequencing platform, according to techniques disclosed herein.

[0043] **FIG. 12A** is a graph of duplex WGS libraries from three starting inputs sequenced at 1-13x coverage, according to techniques disclosed herein.

[0044] **FIG. 12B** is a graph illustrating a duplication rate of the samples of **FIG. 12A**.

[0045] **FIG. 12C** is a graph of the effect of downsampling experiments, according to techniques disclosed herein.

[0046] **FIG.12D** is a graph illustrating that duplex coverage is significantly higher at fixed coverage, according to techniques disclosed herein.

[0047] **FIG. 12E** is a graph illustrating a number of duplex variants found using fgbio versus a decision tree, according to techniques disclosed herein.

[0048] **FIG. 13** is a graph illustrating mutational error rates in mouse PDX samples, according to techniques disclosed herein.

[0049] **FIG. 14** is a bar graph illustrating a number of pre-surgery samples represented in validation experiments, according to techniques disclosed herein.

[0050] **FIG. 15** is a graph illustrating detection of a chemotherapy mutational signature in plasma-free DNA, according to techniques disclosed herein.

[0051] **FIG. 16A** is a bar graph illustrating an apobec signature and measurement, according to techniques disclosed herein.

[0052] **FIG. 16B** is a bar graph illustrating an apobec signature and measurement, according to techniques disclosed herein.

[0053] **FIG. 17** is a computing node according to embodiments of the present disclosure.

DETAILED DESCRIPTION

[0054] Reference will now be made in detail to the exemplary embodiments of the present disclosure, examples of which are illustrated in the accompanying drawings. Wherever possible, the same reference numbers will be used throughout the drawings to refer to the same or like parts.

[0055] The systems, devices, and methods disclosed herein are described in detail by way of examples and with reference to the figures. The examples discussed herein are examples only and are provided to assist in the explanation of the apparatuses, devices, systems, and methods described herein. None of the features or components shown in the drawings or discussed below should be taken as mandatory for any specific implementation of any of these devices, systems, or methods unless specifically designated as mandatory.

[0056] Also, for any methods described, regardless of whether the method is described in conjunction with a flow diagram, it should be understood that unless otherwise specified or required by context, any explicit or implicit ordering of steps performed in the execution of a

method does not imply that those steps must be performed in the order presented but instead may be performed in a different order or in parallel.

[0057] As used herein, the term "exemplary" is used in the sense of "example," rather than "ideal." Moreover, the terms "a" and "an" herein do not denote a limitation of quantity, but rather denote the presence of one or more of the referenced items.

[0058] Cell-free DNA (cfDNA) sequencing for low-burden cancer monitoring is limited by sparsity of circulating tumor DNA (ctDNA), the abundance of genomic material within a plasma sample, and pre-analytical error rates due to library preparation and sequencing errors. Sequencing costs have historically favored the development of deep targeted sequencing approaches for overcoming sparsity in ctDNA detection, but these techniques are limited by the abundance of cfDNA in samples, which imposes a ceiling on the maximal depth of coverage in targeted panels.

[0059] Whole genome sequencing (WGS) is an orthogonal approach to ctDNA-based cancer detection that can overcome the low abundance of cfDNA, supplanting breadth for depth by integrating signal across the entire tumoral mutation landscape. However, the higher cost of WGS limits practical depth of coverage and broad adoption.

[0060] Lower sequencing costs allows for enhanced ctDNA cancer monitoring via WGS. Emerging lower-cost WGS (Ultima Genomics, 1$/Gb) were applied to plasma samples at ~120x coverage. Copy number and single nucleotide variation profiles are comparable between matched Ultima and Illumina datasets, however the deeper WGS coverage enables ctDNA detection at the parts per million range. These lower sequencing costs are further harnessed to implement duplex error-corrected sequencing at the scale of the entire genome, demonstrating a ~3,000x decrease in errors in the plasma of patient-derived xenograft mouse models when compared to raw sequencing reads, and error rates as low as ~10-7 in plasma samples from patients with metastatic melanoma. The highly de-noised plasma WGS is leveraged to undertake cancer monitoring in the more challenging context of low burden melanoma without matched tumor sequencing. In this context, duplex-corrected WGS allowed us to harness known mutational signature patterns for disease monitoring without matched tumors, paving the way for de novo cancer monitoring.

[0061] Deep WGS may be used for ctDNA detection with low-pass sequencing. Low-cost WGS (Ultima Genomics, $1/Gb) may be used to plasma samples at 120x coverage. Copy number and single nucleotide variation profiles were comparable between matched Ultima and Illumina datasets, however the deeper WGS coverage enabled ctDNA detection at the parts per million range. These lower sequencing costs were further harnessed to implement

duplex error-corrected at the scale of the entire genome, demonstrating a ~3000x decrease in errors in the plasma of patient-derived xenograft mouse models when compared to raw sequencing reads, and error rates as low as ~10^{-7} in plasma samples from patients with metastatic melanoma. The highly de-noised plasma WGS was leveraged to undertake cancer monitoring in the more challenging context of low burden melanoma without matched tumor sequencing. In this context, duplex-corrected WGS allowed the harnessing of known mutational signature patterns for disease monitoring without matched tumors, paving the way for de novo cancer monitoring.Sequencing may be done using Ultima Genomics' sequencing platform to sequence $2.6x10^9 \pm 1.4x10^8$ circulating cfDNA reads from 31 plasma samples (n = 8 healthy controls; n = 19 cancer patient samples; n = 4 patient-derived xenograft mouse samples). Deep WGS (~100x) with *in silico* error correction allows ctDNA detection within the part per million range over the sequenced genome. The cost-effective and ultra high-throughput nature of Ultima Genomics sequencing (10^9 reads per run at 1$/Gb) can produce 4.15 ± 3.23 coverage (range 1.02-11.55) duplex-sequencing libraries of cfDNA and achieve error rates as low as $2.7x10^{-7}$. Together, the results of sequencing and running the *in silico* error correction demonstrate the feasibility and utility of deep WGS in clinical samples for ctDNA detection.High-throughput short-read sequencing has revolutionized the liquid biopsy field, and sequencing costs have historically decreased at a rate faster than Moore's Law. However, the decrease in cost has stagnated over the past decade. The emergence of a high-throughput, low-cost sequencing platform would allow for deeper and broader sequencing of samples and patient populations, respectively. To work towards this goal, a comparative analysis of Ultima and Illumina short-read sequencing platforms was performed, demonstrating that these two approaches have comparable tumor-informed analysis capabilities for circulating tumor DNA detection (**FIG. 2A-2F**). Importantly, these tumor-informed approaches allow for ctDNA detection even at the part per million range, thereby making such analysis suitable for minimal residual disease detection. To further expand and demonstrate the utility of the technology, the more challenging problem of ctDNA detection in tumor-agnostic settings, where disease status or tumor origin is unknown, was addressed. For this important clinical context, whole-genome duplex correction was employed to achieve low error rates, allowing us to deconvolve the cell-free DNA mutational compendium into representative mutational signatures to detect ctDNA in the pre-operative setting, without matched tumor sequencing (**FIG. 3A-3H**). Compared to commonly-used off-the-shelf panels, whole-genome analysis has the benefit of sequencing breadth, allowing for the detection of rare tumor-derived mutations that may not be present in targeted panels.

One of ordinary skill in the art may envision that the methods can be harnessed for de novo cancer monitoring in low burden disease scenarios, providing a powerful tool for diagnosing cancer and detecting relapses at the earliest stages, leading to better patient outcomes overall. In addition to de novo cancer detection, the method can be used for cancer screening (*e.g.*, screening for bladder cancer, melanoma, lung cancer, lynch syndrome cancer, BRCA syndrome cancers, based on APOBEC, UV, tobacco, MSI, and BRCA signatures, respectively).

[0062] It is noted that this method is not only useful for de novo detection (e.g. signatures) for tumor monitoring, but also for using tumor informed approaches. In this case, the duplex sequencing will decrease error and provide enhanced signal resolution for detection of a compendia of tumor-confirmed mutations. FIG. 2C illustrates the use of a tumor informed approach, which does not rely on signature analysis.

[0063] In an embodiment, this method is not only useful for monitoring, but also for non-invasive whole genome characterization of mutations in cancer (for example to identify actionable driver mutations or mutations that stratify patients to specific therapies). This is done again via reducing error of various sorts. FIG. 2B illustrates this characterization, which does not rely on signature analysis.

[0064] In an embodiment, this method enables non-invasive detection and discovery of driver mutations in somatic mosaicism. FIG. 2E shows detection of non-malignant mutation, specifically, detection of clonal hematopoiesis mutations.

[0065] **FIG. 1** is a flowchart illustrating an exemplary method for detecting variants in DNA using duplex sequencing and denoising, according to an exemplary embodiment of the present disclosure. For example, an exemplary method **100** (e.g., steps **102-118**) may be performed, in part, by a processor automatically or in response to a request by a user.

[0066] According to one embodiment, the exemplary method **100** for detecting variants may include one or more of the following steps. In step **102**, the method includes extracting DNA from a collection of plasma samples. Plasma can be from any human fluid, including urine, saliva, peritoneal fluid, cerebral spinal fluid, *etc.* The extracted DNA may be cfDNA or genomic DNA.

[0067] Genomic DNA can be extracted using the QiAamp DNA Mini Kit (Qiagen, cat# 563034) and the QiAamp DNA blood Kit (Qiagen, cat# 51104) for tissue and blood samples, respectively, and sheared to 450bp (Covari). In an exemplary experiment, sequencing libraries were prepared on 1µg of DNA using the TruSeq DNA PCR-Free Library Preparation Kit (Illumina), with one additional bead cleanup performed after end-repair and

after adapter ligation. Extracted DNA was quantified using a Qubit 3.0 fluorometer and length analysis was performed using an Agilent Bioanalyzer or High Sensitivity Fragment Analyzer. 2x150bp paired-end sequencing was performed on either a HiSeq X or NovaSeq v1.0 Illumina machine.

[0068] Cell-free DNA can be extracted from plasma using the Magbind cfDNA extraction kit (Omega Biotek, M3298). Manufacturer recommendations for extraction were followed, but elution volume was increased to 35uL and elution time was increased to 20 minutes on a thermomixer at 1,600 rpm (room temperature). Extracted cfDNA was quantified using a Qubit 3.0 fluorometer and length analysis was performed using an Agilent Bioanalyzer or High Sensitivity Fragment Analyzer.

[0069] In step **104**, the method includes preparing a whole genome library with duplex adapters. In some embodiment, the library is prepared by ligating a duplex adapter having a three base pair Unique Molecule Identifier (UMI) and amplifying an amount of DNA with a first PCR to a level that allows for sequencing.

[0070] The duplex adapters contain a three base pair UMI that allows for tracing a top strand of the DNA to a bottom strand of the native DNA molecule. In rare mutation settings such as early detection and minimal residual disease, ctDNA content can fall below 1 in 10,000 concentrations. Therefore, in 1 mL of plasma containing 1,000-10,000 GEs, at most 1 circulating tumor read can be expected to overlap each somatic locus, which is lower than the per-base error rate of high-throughput sequencers (~1 error per 1000 bases). To overcome this issue, deep-targeted sequencing approaches can use UMIs that are incorporated during library preparation for sequencing error correction. While strand-agnostic UMIs can help collapse sequencing errors, UMIs that link forward and reverse DNA strands (i.e. Duplex sequencing) can be used to collapse errors that arise on one strand (such as G>T transversions due to oxidative DNA damage) or during library preparation.

[0071] Extracted cfDNA libraries can be generated in a similar fashion as in Illumina whole genome sequencing, although the full-length adapters are replaced with stubby Y-adapters containing the three UMI bases (IDT Duplex Seq adapters (1080799)).

[0072] A first PCR amplification creates a set of PCR duplicates that increase the amount of DNA allowing for sequencing. These duplicates can be used to remove sequencing errors when two or more molecules with the same UMI are mapped back.

[0073] In an exemplary experiment, six PCR cycles were carried out using indexing primers for input masses above 5ng, and 8 cycles were performed for < 5ng. Libraries were quantified as described above. To enhance duplicate recovery in human samples, 4ng of

prepared libraries was subjected to 6 additional PCR cycles prior to Ultima library conversion. Mouse PDX samples did not undergo additional PCR cycles prior to Ultima library conversion.

[0074] In step **106**, the method includes selecting a subset of the whole genome library and amplifying the subset with a second PCR to increase an amount of PCR duplicates. As with the first PCR amplification, these duplicates can be used to remove sequencing errors when two or more molecules with the same UMI are mapped back.

[0075] In step **108**, the method may include sequencing a series of reads from the amplified subset.

[0076] *Illumina sequencing*: Illumina sequencing libraries were sequenced on a HiSeq X or NovaSeq1.0 using 2x150 paired-end sequencing.

[0077] *Ultima sequencing*: Illumina sequencing libraries were sent to Ultima Genomics (Newark, CA) for library conversion and sequencing.

[0078] In step **110**, the method includes aligning the reads to a host genome (*e.g.*, the human genome) and denoising the duplex reads, using original sequencing reads and collapsed read information.

[0079] FastQ reads were adapter and UMI trimmed using cutadapt (version X). Trimmed reads were then aligned to the human genome (version hg38) using bwa mem (with parameters -K 100000000 -p -v3 -t 16 -Y). Trimmed UMI's were added to the alignment files as an additional RX tag. Single-strand and duplex correction was carried out using the fgbio suite of tools (version 2.0). For single-strand error correction, reads were grouped by UMI (fgbio GroupReadsByUmi -s edit) and consensus calls were performed (fgbio CallMolecularConsensusReads --min-reads 2). Resulting error-collapsed fastQs were realigned to the human genome using bwa mem. For duplex error correction, single-strand consensus sequencing was performed independently on top-strand-mapping and bottom-strand-mapping reads. Bottom-strand mapping reads were subsequently reverse complemented and merged with top-strand mapping reads. Reads were then re-grouped by UMI and error correction was performed to obtain duplex corrected reads. Uncorrected reads that belonged to a duplex family were processed to measure the following read-specific features:

[0080] 1. Base pair at the variant position

[0081] 2. Edit distance of the read to the reference

[0082] 3. Total number of single-nucleotide variants on the read

[0083] 4. Read mapping quality

[0084] 5. The position along the read from the extremities of the DNA fragment

[0085] In step **112**, the method includes extracting a variant from the duplex reads. The type of variant is defined by a mutation, for example, a C base mutated into T (represented as C>T) and the base pairs adjacent to the mutated base. For example: A[A>T]A is one type of variant. A[A>C]A is another type of variant, as is A[A>G]A, A[C>A]A, etc. There are 96 types of variant.

[0086] Reads are filtered based on whether 1) all reads in a duplex family carried the same variant; 2) the edit distance was lower than 2; 3) the total number of single-nucleotide variants on the read was below 10; 4) the mapping quality was the highest possible value (60 for bwa mem); 5) the position along the read was greater or equal to 10.

[0087] In step **114**, the method includes comparing the extracted variant to a collection of cancer specific variant signatures. Certain cancers have very well-defined variant types, *i.e.*, cancer specific mutational signatures. Melanoma, for example, is a cancer with a distinct signature that is related to the skin's exposure to UV rays. Some lung cancers will show a signature associated with exposure to tobacco. This signature matching process is shown by **FIG. 3E**. The fraction measurement estimation represented along the y-axis is found by using copy-number based tumor fraction estimation.

[0088] Copy number analysis was performed using ichorCNA (version). Tumor fractions were estimated after correcting for library and sequencing artifacts via a panel of normals from cancer-free controls (CTRL-01 to CTRL-05) sequenced on the same instrument as the sample.

[0089] In low tumor burden settings, not all somatic mutations from the tumor are represented in the cell-free DNA sequencing pool, and any genomic locus is expected to be covered by at most one circulating tumor DNA read. Therefore, read-based TF estimation frameworks, and not locus-based TF estimations, are necessary to accurately quantify ctDNA content. Genome-wide mutations from the sequencing reads may be integrated and summarized as a weighted sum of single-base substitution (SBS) reference mutational signatures. Expectedly, in a re-analysis of publicly available PCAWG melanoma datasets, it was found that the UV-associated SBS7 mutational signature was most abundant in melanoma WGS datasets, and that clock-like signatures 1A/B weakly correlated with age of patients at tissue collection (spearman's $\rho = 0.26$, p-value = 0.0071, **FIGs. 6A-6B**). The trinucleotide contexts of cfDNA variants were explored through mutation signature analysis at each level of denoising (UMI-agnostic, single-stranded and duplex), in order to investigate the potential sources conferring mutations in these samples. Cosine similarities between the

UV-associated SBS7 signature and high burden samples were highest after duplex correction (mean cosine similarities to SBS7 of 0.967±0.02 and 0.365±0.03 between duplex corrected and uncorrected samples, p-value = $1.6x10^{-4}$ (Wilcoxon rank sum test)), **FIGs. 7A-7B**). Similar improvements were found when measuring cosine similarities between clock-like signatures and cancer-free controls, highlighting the importance of duplex correction for accurate signature analysis (**FIGs. 7A-7B**).

[0090] Given the ability of de novo mutation identification in error corrected cfDNA WGS to deliver profiles matching SBS7 and clock-like signatures for identifying melanoma and age-associated circulating DNA fragments, respectively, a tumor-agnostic approach for ctDNA detection was developed based on mutational patterns. As a first step, SBS mutational signatures are deconvolved from plasma ctDNA mixtures using a non-negative maximum likelihood model, and a tumor fraction is estimated by taking the weight of the tumor-associated SBS signature and normalizing by total number of mutations and depth of sequencing. Second, a signature score is calculated in order to determine whether the cancer-associated SBS signatures better explain the observed mutation profiles compared to a random permutation of the cancer-associated motifs. To analytically validate this approach, an *in silico* mixing study was conducted, combining duplex-denoised reads from two high burden ctDNA samples (MEL-12.A and MEL-12.B) and a cancer-free control (CTRL-06) at 10x sequencing depth (after duplex consensus), in varying proportions (expected tumor fractions from 0 to 1%). As a result, estimated tumor fractions were readily detectable at expected tumor fractions of 10^{-4} (receiving operating characteristic area under the curve 0.90), with signature scores highly specific for melanoma at 10^{-3} dilutions (**FIG. 3F**).

[0091] In an embodiment, a signature-based ctDNA detection platform for pre-operative ctDNA detection (i.e. tumor-agnostic ctDNA detection) was applied. Plasma samples were sequenced from four patients with stage III melanoma, three cancer-free controls, and one treatment-unresponsive patient (5 separate time points) with stage IV melanoma. Notably, signature scores for ctDNA detection showed perfect separation between cancer-free controls and samples from melanoma patients (**FIGs. 3G-3H**), while copy-number based analysis could not resolve these two groups (**FIG. 8**).

[0092] Tumor genotyping can be performed via cfDNA & normal tissue sequencing when the tumor burden in the plasma is high (>10%, sources). Mutect2 can be used with the normal tissue. A quality threshold was established, and only SNVs are kept. Then, four blacklists are applied to create a final tumor panel. (encode, gnomad, local blacklist, centromeres)

[0093] Read counting was performed in 1Mbp bins using hmmcopy (excluding duplicates and reads with a mapping quality below 60). Read counts were adjusted for mappability and GC content using hmmcopy. Separate panels of normals were created for Illumina and Ultima datasets, respectively, using cancer-free controls (CTRL-01 to CTRL-05; n = 5 per sequencing instrument). Tumor fraction estimates were obtained using ichorCNA (version). For plotting purposes (Figure 1X, Supplemental Figure Y), corrected log2 read counts outputted by ichorCNA were used. Bins marked by ichorCNA as copy gains, amplifications and high-level amplifications were marked and colored as chromosome gains (pink). Bins marked as homozygous deletion states and hemizygous deletions were marked and colored as chromosome losses (blue). Copy neutral regions were marked as neutral (black). Bins with corrected log2 read counts between -0.05 and 0.05 were marked as neutral (black) as well.

[0094] In step **116**, the method includes determining a cancer status based on a match between variant signatures.

[0095] Variants detected using the denoising method described above in were used. Variants with allele frequencies greater than 30% were presumed to be germline mutations and were discarded. Remaining reads were aggregated, and the frequencies of the variants in their trinucleotide context were calculated. These trinucleotide variant frequencies were compared to the trinucleotide variant frequencies of publicly available references for different biological processes. In this context, given that processed samples were from cancer-free controls and melanoma patients, it was assumed that the sample's trinucleotide variant frequencies would be a combination of aging-related trinucleotide frequencies and UV-damage associated trinucleotide frequencies. The sample's frequencies were fit to the references using a non-negative maximum likelihood method. To remove false positives, a permutation test was performed. This test involved randomly changing the trinucleotide frequencies of the UV-associated (melanoma) reference signature and performing the non-negative maximum likelihood fit. If the sample showed a stronger fit to the randomly-permuted frequencies than to the original one, it was deemed to be a false positive. This exercise was repeated 10,000 times to obtain a signature score. If the sample had a signature score below 0.001, it was deemed acceptable. Samples above this threshold were deemed cancer-negative.

[0096] In some embodiments, the method may include exhaustive WGS. In addition to WGS, embodiments of the present disclosure may use less exhaustive sequencing methods such as whole exome sequencing (WES) or SNP genotyping. Various enrichment modalities may be employed, including but not limited to: exome enrichment, targeted gene enrichment,

and/or specific mutation enrichment. For example, exome enrichment may comprise whole exome sequencing. Targeted gene enrichment may include sequencing entire genes, including one or more of introns, exons, and/or coding sequences. A specific mutation enrichment may target a specific position of the genome, including, *e.g.*, an exon, an intron, and/or some other user-defined position. Technologies to accomplish enrichment of at least one of the aforementioned regions are typically hybridization based or primer based. Such examples include: targeted variant sequencing; targeted gene sequencing; whole exome sequencing; targeted PCR; nested PCR; and/or linear PCR.

[0097] In cases where a limited input is used, such as in cell-free DNA, it is possible to exhaustively sequence the sample (*i.e.*, sequence every single molecule available). Accordingly, some embodiments of the present disclosure may include exhaustive whole exome sequencing, exhaustive cfDNA sequencing, and/or exhaustive targeted sequencing.

[0098] **FIGs. 2A-2F** depict ultralow ctDNA detection requiring deep sequencing coverage and low error rates.

[0099] **FIG. 2A** is a collection of graphs showing a simulated sequencing coverage. Simulation analysis shows that lower error rates and high sequencing coverage are required for accurate ctDNA detection when tumor fractions are at or below 10^{-5}.

[00100] Simulations for **FIG. 2A** were performed assuming a tumor-mutational compendium of 10,000 SNVs at different error rates (10-3, 10-4 and 10-5), coverages (1, 10 and 100) and tumor fractions (0, 10-6, 10-5). For each of the 50,000 SNV mutations, coverage was simulated using a poisson distribution. Each simulated sequenced base pair was classified as either ctDNA or cfDNA according to the tumor fraction, and errors misclassified as ctDNA were determined according to the error rate. Estimated tumor fractions were calculated by summing the ctDNA molecules and the errors, and dividing by the total base pairs simulated.

[00101] **FIG. 2B** is a pre-analytical workflow for cfDNA library preparation. The workflow, similar to that of the embodiment shown in **FIG. 1**, comprises obtaining plasma, from which cfDNA is extracted. The double stranded DNA library is prepared for sequencing, which is done using the Illumina sequencing method or Ultima library conversion and subsequent Ultima sequencing.

[00102] **FIG. 2C** is a graph comparing sequencing depth (genome equivalents) of matched Illumina and Ultima datasets, across 15 matched cfDNA samples.

[00103] **FIG. 2D** is a comparison of normalized read coverage for Illumina (top) and Ultima (bottom) matched cfDNA samples (chromosomes).

[00104] **FIG. 2E** is a comparison of copy number variations (CNV) tumor fraction and single-nucleotide variants (SNV) tumor fraction using Illumina and Ultima datasets. On the left graph, the CNV tumor fraction estimation measured with Illumina or Ultima sequencing is shown in matched samples using ichorCNA. Matched cancer-free controls were used to create a panel of normal prior to tumor fraction estimation.

[00105] On the right graph of **FIG. 2E**, single nucleotide variant-based tumor fraction estimation measured with Illumina or Ultima sequencing is shown. Somatic SNVs were identified through matched tumor-normal sequencing. Two samples without tumor sequencing and with low ctDNA fraction (e.g., less than 5% measured through CNV analysis) were omitted.

[00106] **FIG. 2F** depicts an expected tumor fraction score with and without error suppression. An *in silico* mixing study of metastatic melanoma sample MEL-01 with cancer-free control CTRL-05 (50 replicates per tumor fraction, 80x coverage per replicate) show the effect with (red) and without (blue) tumor-informed analytic denoising applied using Ultima-specific quality filtering.

[00107] **FIGs. 3A-3H** depict duplex correction allowing ctDNA without tumor sequencing. **FIG. 3A** depicts error rates in mouse and human DNA among duplex sequencing, single strand sequencing, and uncorrected groups. The graph on the left shows error rates for duplex WGS sequencing on mouse PDX samples (n=3). Open circles in the graph on the left represent samples for which no sequencing errors were detected. The graph on the right represents duplex WGS sequencing in patient sample MEL-12.D intersected with tumor mutation profiles of 107 melanoma patients retrieved from the Pan Cancer Analysis of Whole Genome Consortium. Base changes matching the somatic mutation of the tumor were considered errors (after removing germline and somatic mutations from the matched patient data).

[00108] To first test the accuracy of duplex error correction, duplex libraries were prepared using cfDNA obtained from the plasma of mice with patient-derived xenografts (n=4, NOD/ShiLtJ species; n = 1 lung cancer; n = 3 diffuse large B cell lymphoma). Tumor fractions, defined as the fraction of reads uniquely mapping to the human genome, were 0.4%, 40%, 73% and 96%. To estimate the error rate of the duplex libraries, mutation levels were investigated at well characterized homozygous variant sites for NOD/ShiLtJ mice in the three samples with elevated mouse-mapped reads. Overall, only two bases out of over 4.2×10^6 total bases were sequenced that were inconsistent with the known genotype of the mice for an error rate of 4.75×10^{-7} (**FIG. 3A**). These results are consistent with a previous

report employing whole genome duplex sequencing (Abascal et al, 2021 reports error rates of $2x10^{-7}$ using similar protocols).

[00109] **FIG. 3B** is a series of graphs comparing variant allele frequencies calculated using unfiltered sequencing reads. Variant allele frequencies are shown in positions where a variant was found using uncorrected reads (left column) and in duplex corrected reads (right column). Top and bottom rows are representative examples for cancer-free and high-burden patient samples, respectively.

[00110] **FIG. 3C** is a graph comparing the model allele fraction of a patient with progressive disease (samples MEL-12.A-E) in duplex corrected positions (allele fractions below 30% only) and copy-number based tumor fraction estimations.

[00111] **FIG. 3D** and 3E illustrate an exemplary method of signature matching between sequencing reads. The signature 7 reference of **FIG. 3D** is a publicly available signature associated with UV exposure (*i.e.*, a melanoma specific signature). The signatures of **FIG. 4E** show uncorrected, single- strand correction, and duplex correction of a control signature and a melanoma patient with a MEL-12 D signature. Each bar on the signature represents a specific trinucleotide mutation (*i.e.*, there are 96 bars) and the y-axis shows the relative proportion of trinucleotide mutation. The duplex corrected MEL-12 D signature is matched to the reference signature, and the cancer-like signature is only apparent in the cancer patient after duplex correction.

[00112] **FIG. 3F** is a graph illustrating a signature score and ctDNA detection of an *in silico* mixing study of metastatic melanoma samples MEL-12.A/B with cancer-free control CTRL-06 (10 replicates per tumor fraction, 10x coverage per replicate).

[00113] In the top row, the signature score is used to estimate the contribution of signature SBS& (melanoma UV associated) in the decomposition of a sample's trinucleotide frequencies into reference signatures.

[00114] In the bottom row, ctDNA detection by expected tumor fraction. Z-scores estimation was used to calculate mutation signature SBS7 detection in comparison to detection in TF=0 replicates. Ground truth variants originating from either the high-burden sample MEL-12.A/B. or the cancer-free sample CTRL-06 are shown in blue (full circle: MEL-12.A/B; open circle: CTRL-06). Error bars represent the standard deviation in the number of variants per replicate at a given expected tumor fraction.

[00115] **FIG. 3G** is a graph illustrating a series of signature scores of melanoma signature 7 in plasma cfDNA samples using duplex WGS (n=9 melanoma samples; n=3 controls). Samples in red are from patient MEL-12 with stage IV melanoma at different time points in

their clinical course. Samples in blue each represent a separate patient (MEL-08 to MEL-11). Samples in pink represent control samples.

[00116] **FIG. 3H** is a graph illustrating estimated tumor fraction of samples with elevated signature scores. The X-axis depicts the clinical timepoint for each patient sample. Tumor fractions were estimated by multiplying the number of single nucleotide variants found in duplex corrected reads by the weight of signature 7 after reference signature decomposition and normalization by depth of coverage. An assumed 10,000 SNV tumor mutational profile was assumed for tumor fraction estimation.

[00117] **FIG. 4** depicts cfDNA fragment lengths in single-end sequencing datasets matched with paired-end sequencing. Fragment lengths are accurately recovered between single-end Ultima reads when compared to paired-end Illumina sequencing for cfDNA molecules below 200 base pairs.

[00118] **FIGs. 5A-5C** depict the effective of artifact blacklisting on a single nucleotide variant detection.

[00119] **FIG. 5A** is a graph illustrating the UG specific blacklist. The UG specific blacklist includes regions with low GC content, tandem repeats, regions with poor mappability, regions with high coverage variability and regions with homopolymers greater than 10 base pairs.

[00120] **FIG. 5B** is an illustration of overlap between the UG blacklist and other low confidence regions. Other low confidence regions include centromeres, simple repeats, regions that encode blacklist, and gnomad regions with AF value greater than 0.001.

[00121] **FIG. 5C** is a graph illustrating the overlap between melanoma tumor tissue SNVs and low confidence regions. The effects of blacklists on the recovery of somatic single nucleotide variants (SNVs) are shown in 107 melanoma tissue samples obtained from the Pan Cancer Analysis of Whole Genomes consortium.

[00122] **FIGs. 6A-6B** depict cosine similarities in high burden and cancer-free samples for clock-like and UV-associated signatures SBS1B and SBS7, respectively.

[00123] **FIG. 6A** is a heatmap of cosine similarities in cancer-free samples and high-burden ctDNA samples. The heatmap of cosine similarities in duplex-corrected, single strand corrected and uncorrected reads from cancer-free samples (n=3) or high burden ctDNA samples (n=5, all from patient MEL-12 with stage IVB melanoma). The x-axis is ordered alpha-numerically, without hierarchical clustering

[00124] **FIG. 6B** is a boxplot of cosine similarities for three correction methods in the same cancer-free samples and high-burden ctDNA samples, according to techniques disclosed herein.

[00125] **FIGs. 7A-7B** depict the re-analysis of 107 melanoma mutational signatures from the Pan-Cancer Analysis of Whole Genomes consortium.

[00126] **FIG. 7A** is a graph showing the signature fraction of a number of variant signatures. Deconvolution of duplex-corrected mutations into representative mutational signatures was performed using a non-negative maximum likelihood model. Boxplots are ordered in increasing order of the median for each signature.

[00127] **FIG. 7B** is a correlation plot between age at cancer diagnosis and the number of clock-like mutations attributed to SBS1A and SBS1B. The number of mutations was obtained by multiplying the weights of SBS1A and SBS1B by the total number of mutations found after duplex correction.

[00128] **FIG. 8** depicts tumor-agnostic copy-number based tumor fraction estimation in cancer-free control samples (n=3) and pre-surgery melanoma plasma (n=4).

[00129] In another embodiment, whole genome sequencing may occur without duplex, to reach an SNV-based tumor fraction estimation

[00130] SNV-based tumor fraction estimation was carried out by counting cell-free DNA reads with matching tumor-specific somatic mutations (mutation calling pipeline described below). To limit the effect of problematic regions of the genome, a platform-specific blacklist was built. For Illumina sequencing, regions identified in the ENCODE blacklist (source), centromeres (source), simple repeat regions (source) and positions with high mutation rates (GNOMAD, AF>0.001, source) were not considered. For Ultima sequencing, Ultima-specific low-confidence regions composed of homopolymers, AT-rich regions, tandem repeats and regions with poor mappability and high coverage variability were additionally excluded.

[00131] To limit the effect of sequencing errors, custom scripts were used for platform-specific denoising. Illumina alignment files were filtered to contain read pairs overlapping somatic mutation positions. Paired-end reads were filtered for X, Y, Z and were only kept if both R1 and R2 carried the somatic mutation or the reference base pair. Tumor fractions were estimated by dividing the number of filtered reads containing the somatic mutation by the total number of filtered reads.

[00132] Ultima alignment files were subset to reads overlapping with somatic mutation positions. Reads were filtered by X, Y, Z. Tumor fractions were estimated by dividing the

number of filtered reads containing the somatic mutation by the total number of filtered reads.

[00133] *SNV model training sets and feature space*

[00134] Training sets were obtained from plasma enriched for ctDNA SNV fragments (true label) from specific melanoma tumors and cfDNA SNV reads (false label) from healthy controls without known cancer as listed in sup tab xx. Candidate reads were extracted from custom denoised alignment files. For true label sets, patients with high burden metastatic disease were used and only reads which represented matched tumor variants were retained.

[00135] A custom deep-learning model is used for signal to noise enhancement, similar to previous work (Widman et al, 2022), and effectively categorized candidate SNV reads. Candidate SNV reads were extracted using pysam (v0.15.2). Additionally, compelling regional and sequencing tech specific features were encoded as input to the deep learning model architecture with a custom python (v3.6.8) script. Two separate input structures are described below, corresponding to each component of the ensemble model.

[00136] For the MLP, a tabular set of feature values is provided as an input.

[00137] The feature selection for this was performed on SNV reads post filtering in both the true and false label settings. Specific features and their corresponding single variable AUC performance is described in sup tab xx. As highlighted in previous work (Widman et al, 2022), tissue-specific transcriptional features aid in defining the likelihood for observing somatic mutations in a genomic region. Local tumor mutation density is categorized by quantifying WGS SNV mutation calls from the PCAWG database (edge ref 81) and the total number of SNV mutations are counted from available melanoma derived tumor samples. Additionally, local histone CHiP-Seq marks and tissue specific bulk RNA expression values were reported as standard RPKM values from primary tissue alignments in ENCODE (edge ref 95). Regional DNase peaks (lifted to GRCh38) were also included, which were obtained from narrowpeak files as reported in ENCODE (edge ref 95,96). Melanoma specific ATAC peak calls as reported in TCGA (edge ref 82) were also included.

[00138] Since the deep learning model is designed to operate on a read level compendium, values for the features defined above were computed using a sliding window around each candidate read. The optimal length for this sliding window was defined in previous work (Widman et al, 2022). Additionally, regional chromatin annotation tracks (ChromHMM - lifted to GRCh38) (edge ref 83) were obtained from ENCODE. Hi-C SNIPER(edge ref 97) bed files were used to extract HI-C compartment information. Lastly, regional features for

replication timing and mean expression values (lifted to GRCh38) were pulled from previous literature (edge ref 37).

[00139] In addition, Ultima specific read level features were included. These include the following:

- X-FC1 - number of features (SNPs) on the same read
- X-FC2 - number of features (SNPs) on the same read that passed the filter (matching the reference for -+5 bases)
- X-FLAGS – propagated from the bam file flag.
- ed_1- edit (Levenshtein) distance of the read from the reference, before SNV of interest.
- ed_2- edit (Levenshtein) distance of the read from the reference, after SNV of interest.

[00140] Next for the CNN a one-hot encoded tensor structure of the candidate read was used, similar to previous work (Widman et al, 2022). Each read is encoded with a variant (sequencing artifact/noise from healthy controls or somatic mutation from high burden tumor plasma sample). The encoded tensor has an image-like structure with a shape of 12x240. The rows correspond to one hot encoded nucleotides (N,A,C,T,G) corresponding to the reference and the read. The penultimate row dimension is used to mark the position along the read highlighting the SNV of interest. Lastly, the absence/presence (0/1) of a cycle skip (as defined by Ultima) is encoded along the last row dimension to add further relevance to tri-nucleotide context of the SNV of interest. The columns correspond to individual nucleotides along the length of the read. While reads have a maximum length of 200, the extra 40 base pairs are padded with the reference genome thereby adding additional relevant contextual information.

[00141] SNV model design and training

[00142] The deep-learning model has an ensemble structure and consists of two major components - a regional/read specific multi layer perceptron (MLP) and a sequence based convolutional neural network (CNN), whose weight matrices are jointly learnt.

[00143] The MLP which takes a feature matrix as input consists of a linear stack of four dense blocks. Each block is defined as consisting of a fully connected layer with a ReLU activation. Furthermore, for the purpose of regularization the input to each fully connected layer is batch normalized and the output is passed through a dropout layer.

[00144] The CNN consists of four one dimensional convolution layers with non-linear ReLU activations, which extract sequential information at different spatial resolutions. Moreover, as in classical deep learning frameworks, each convolution layer (post nonlinear activation) is followed by a max pooling layer. The output is then passed through a stack of 3 dense blocks as defined above.

[00145] Subsequently, the latent output of both the MLP and CNN is then concatenated and passed through a single dense block. Finally, a probability score between 0 (sequencing noise) and 1 (true somatic mutation) is obtained by using a single sigmoid-activated fully-connected layer. This probability score reflects the model's estimate on whether a candidate SNV mutation present in the encoded read is likely from signal or noise. The ensemble model is built in Keras (v.2.3.0) with a Tensorflow base (1.14.0).

[00146] To train the ensemble model, the objective function defined as a binary cross entropy loss is minimized. Performance metrics were reported within balanced sets.

[00147] *UMI Correction improves insertion-deletion mutation (indel) detection accuracy in Ultima sequencing datasets*

[00148] UMIs add a unique barcode to each DNA molecule. During PCR, the barcode tag (and DNA molecule) is duplicated multiple times. PCR duplicates can be thereby identified using the UMI. Identified duplicates can be used to correct sequencing errors, as it is unlikely that the same error will occur on two PCR duplicates. It should be noted that the Ultima flow-based sequencing is prone to homopolymer size errors, which are interpreted as false indel. **FIG. 9A** is a graph illustrating a homopolymer size between two PCR duplicates, whereas **FIG. 9B** is a graph illustrating a homopolymer size between a read and an aligned reference. For reference, **FIG. 9C** is a graph illustrating frequency of homopolymer size across the human genome, according to techniques disclosed herein. To further illustrate the increased accuracy of UMI correction, **FIG. 9D** is a graph illustrating insertion-deletion mutations indel calling accuracy by PCR duplicate family sizes, according to techniques disclosed herein.

[00149] *UMI ligated reads allow for the detection of error robust trinucleotide motifs in Ultima sequencing datasets*

[00150] UMIs can also be used to find error-robust single nucleotide variants. These variants generally fall in cycle shift motifs, which are specific trinucleotides that Ultima has determined to be robust to errors. Details on cycle shifts are shown in **FIG. 10**, a graph illustrating a single nucleotide variant analysis of matched Ultima and Illumina sequencing datasets. **FIGs. 11A-B** show that flow-based sequencing provides predictable error-robust

motifs. **FIG. 11A** is a flow chart of a sequencing process providing predictable, error-robust motifs, and **FIG. 11B** is a graph of error rate by sequencing platform, according to techniques disclosed herein.

[00151] *Wetlab improvements and the development of a novel machine learning classifier for duplex variant detection*

[00152] In some embodiments of the present disclosure, molecular and computational improvements were made to the methodology to improve yield of duplex molecules. The molecular improvement is shown by a more efficient bottlenecking of DNA molecules (**FIG. 12A-B**). **FIG. 12A** is a graph of duplex WGS libraries from three starting inputs sequenced at 1-13x coverage. Duplex correction was applied, and the yield of duplex recovery (depth of duplex-only coverage by total sequenced coverage) was measured, as shown in the graph. **FIG. 12B** illustrates the duplication rate of the samples of **FIG. 12A** at n = 3.

[00153] In **FIG. 12C**, a downsampling experiment shows that improved bottlenecking achieves higher duplex coverage at a faster rate than other embodiments. This is further illustrated by **FIG. 12D**, which is a a graph illustrating that duplex coverage is significantly higher at fixed coverage. **FIG. 12E** is a graph illustrating a number of duplex variants found using the duplex method (fgbio) versus a decision tree. **FIG. 13** furthers this illustration with mutational error rates in mouse PDX samples (with N = 3 per condition). This data illustrates the expanded applicability of the processes of the disclosed embodiments, including application to melanoma, stage III, with a baseline timepoint.

[00154] Some embodiments of the present disclosure produce more samples, including pre-surgery samples as shown in **FIG. 14**. Plasma cell-free DNA was obtained from patients with bladder cancer who may or may not have received chemotherapy. **FIG. 15** shows that embodiments of the present disclosure detect a chemotherapy mutational signature in most samples that may have received chemotherapy, and specifically illustrates the application to bladder cancer, and the detection of an "APOBEC" signature and chemotherapy. In samples who never received chemo (green) or cancer-free controls (blue), the chemotherapy signal is not measured.

[00155] Bladder cancer typically shows the APOBEC mutational signature. This signature can also be detected in the plasma cell-free DNA, as shown in **FIGs. 16A-B**. Darker bars represent the APOBEC signature of tumors, and the lighter bars represent the APOBEC measurement in cfDNA.

[00156] **FIG. 17** is a schematic of an example of a computing node. Computing node **10** is only one example of a suitable computing node and is not intended to suggest any limitation

as to the scope of use or functionality of embodiments described herein. Regardless, computing node **10** is capable of being implemented and/or performing any of the functionality set forth hereinabove.

[00157] In computing node **10** there is a computer system/server **12**, which is operational with numerous other general purpose or special purpose computing system environments or configurations. Examples of well-known computing systems, environments, and/or configurations that may be suitable for use with computer system/server **12** include, but are not limited to, personal computer systems, server computer systems, thin clients, thick clients, handheld or laptop devices, multiprocessor systems, microprocessor-based systems, set top boxes, programmable consumer electronics, network PCs, minicomputer systems, mainframe computer systems, and distributed computing environments that include any of the above systems or devices, and the like.

[00158] Computer system/server **12** may be described in the general context of computer system-executable instructions, such as program modules, being executed by a computer system. Generally, program modules may include routines, programs, objects, components, logic, data structures, and so on that perform particular tasks or implement particular abstract data types. Computer system/server **12** may be practiced in distributed computing environments where tasks are performed by remote processing devices that are linked through a communications network. In a distributed computing environment, program modules may be located in both local and remote computer system storage media including memory storage devices.

[00159] As shown in **FIG. 17**, computer system/server **12** in computing node **10** is shown in the form of a general-purpose computing device. The components of computer system/server **12** may include, but are not limited to, one or more processors or processing units **16**, a system memory **28**, and a bus **18** that couples various system components including system memory **28** to processor **16**.

[00160] Bus **18** represents one or more of any of several types of bus structures, including a memory bus or memory controller, a peripheral bus, an accelerated graphics port, and a processor or local bus using any of a variety of bus architectures. By way of example, and not limitation, such architectures include Industry Standard Architecture (ISA) bus, Micro Channel Architecture (MCA) bus, Enhanced ISA (EISA) bus, Video Electronics Standards Association (VESA) local bus, Peripheral Component Interconnect (PCI) bus, Peripheral Component Interconnect Express (PCIe), and Advanced Microcontroller Bus Architecture (AMBA).

[00161] Computer system/server **12** typically includes a variety of computer system readable media. Such media may be any available media that is accessible by computer system/server **12**, and it includes both volatile and non-volatile media, removable and non-removable media.

[00162] System memory **28** can include computer system readable media in the form of volatile memory, such as random access memory (RAM) **30** and/or cache memory **32**. Algorithm Computer system/server **12** may further include other removable/non-removable, volatile/non-volatile computer system storage media. By way of example only, storage system **34** can be provided for reading from and writing to a non-removable, non-volatile magnetic media (not shown and typically called a "hard drive"). Although not shown, a magnetic disk drive for reading from and writing to a removable, non-volatile magnetic disk (*e.g.*, a "floppy disk"), and an optical disk drive for reading from or writing to a removable, non-volatile optical disk such as a CD-ROM, DVD-ROM or other optical media can be provided. In such instances, each can be connected to bus **18** by one or more data media interfaces. As will be further depicted and described below, memory **28** may include at least one program product having a set (*e.g.*, at least one) of program modules that are configured to carry out the functions of embodiments of the disclosure.

[00163] Program/utility **40**, having a set (at least one) of program modules **42**, may be stored in memory **28** by way of example, and not limitation, as well as an operating system, one or more application programs, other program modules, and program data. Each of the operating system, one or more application programs, other program modules, and program data or some combination thereof, may include an implementation of a networking environment. Program modules **42** generally carry out the functions and/or methodologies of embodiments as described herein.

[00164] Computer system/server **12** may also communicate with one or more external devices **14** such as a keyboard, a pointing device, a display **24**, etc.; one or more devices that enable a user to interact with computer system/server **12**; and/or any devices (*e.g.*, network card, modem, etc.) that enable computer system/server **12** to communicate with one or more other computing devices. Such communication can occur via Input/Output (I/O) interfaces **22**. Still yet, computer system/server **12** can communicate with one or more networks such as a local area network (LAN), a general wide area network (WAN), and/or a public network (*e.g.*, the Internet) via network adapter **20**. As depicted, network adapter **20** communicates with the other components of computer system/server **12** via bus **18**. It should be understood that although not shown, other hardware and/or software components could be used in

conjunction with computer system/server **12**. Examples, include, but are not limited to: microcode, device drivers, redundant processing units, external disk drive arrays, RAID systems, tape drives, and data archival storage systems, etc.

[00165] In various embodiments, a learning system is provided. In some embodiments, a feature vector is provided to a learning system. Based on the input features, the learning system generates one or more outputs. In some embodiments, the output of the learning system is a feature vector. In some embodiments, the learning system comprises an SVM. In other embodiments, the learning system comprises an artificial neural network. In some embodiments, the learning system is pre-trained using training data. In some embodiments training data is retrospective data. In some embodiments, the retrospective data is stored in a data store. In some embodiments, the learning system may be additionally trained through manual curation of previously generated outputs.

[00166] In some embodiments, the learning system, is a trained classifier. In some embodiments, the trained classifier is a random decision forest. However, it will be appreciated that a variety of other classifiers are suitable for use according to the present disclosure, including linear classifiers, support vector machines (SVM), or neural networks such as recurrent neural networks (RNN).

[00167] Suitable artificial neural networks include but are not limited to a feedforward neural network, a radial basis function network, a self-organizing map, learning vector quantization, a recurrent neural network, a Hopfield network, a Boltzmann machine, an echo state network, long short term memory, a bi-directional recurrent neural network, a hierarchical recurrent neural network, a stochastic neural network, a modular neural network, an associative neural network, a deep neural network, a deep belief network, a convolutional neural networks, a convolutional deep belief network, a large memory storage and retrieval neural network, a deep Boltzmann machine, a deep stacking network, a tensor deep stacking network, a spike and slab restricted Boltzmann machine, a compound hierarchical-deep model, a deep coding network, a multilayer kernel machine, or a deep Q-network.

[00168] The present disclosure may be embodied as a system, a method, and/or a computer program product. The computer program product may include a computer readable storage medium (or media) having computer readable program instructions thereon for causing a processor to carry out aspects of the present disclosure.

[00169] The computer readable storage medium can be a tangible device that can retain and store instructions for use by an instruction execution device. The computer readable storage medium may be, for example, but is not limited to, an electronic storage device, a magnetic

storage device, an optical storage device, an electromagnetic storage device, a semiconductor storage device, or any suitable combination of the foregoing. A non-exhaustive list of more specific examples of the computer readable storage medium includes the following: a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), a static random access memory (SRAM), a portable compact disc read-only memory (CD-ROM), a digital versatile disk (DVD), a memory stick, a floppy disk, a mechanically encoded device such as punch-cards or raised structures in a groove having instructions recorded thereon, and any suitable combination of the foregoing. A computer readable storage medium, as used herein, is not to be construed as being transitory signals per se, such as radio waves or other freely propagating electromagnetic waves, electromagnetic waves propagating through a waveguide or other transmission media (*e.g.*, light pulses passing through a fiber-optic cable), or electrical signals transmitted through a wire.

[00170] Computer readable program instructions described herein can be downloaded to respective computing/processing devices from a computer readable storage medium or to an external computer or external storage device via a network, for example, the Internet, a local area network, a wide area network and/or a wireless network. The network may comprise copper transmission cables, optical transmission fibers, wireless transmission, routers, firewalls, switches, gateway computers and/or edge servers. A network adapter card or network interface in each computing/processing device receives computer readable program instructions from the network and forwards the computer readable program instructions for storage in a computer readable storage medium within the respective computing/processing device.

[00171] Computer readable program instructions for carrying out operations of the present disclosure may be assembler instructions, instruction-set-architecture (ISA) instructions, machine instructions, machine dependent instructions, microcode, firmware instructions, state-setting data, or either source code or object code written in any combination of one or more programming languages, including an object oriented programming language such as Smalltalk, C++ or the like, and conventional procedural programming languages, such as the "C" programming language or similar programming languages. The computer readable program instructions may execute entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network,

including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider). In some embodiments, electronic circuitry including, for example, programmable logic circuitry, field-programmable gate arrays (FPGA), or programmable logic arrays (PLA) may execute the computer readable program instructions by utilizing state information of the computer readable program instructions to personalize the electronic circuitry, in order to perform aspects of the present disclosure.

[00172] Aspects of the present disclosure are described herein with reference to flowchart illustrations and/or block diagrams of methods, apparatus (systems), and computer program products according to embodiments of the disclosure. It will be understood that each block of the flowchart illustrations and/or block diagrams, and combinations of blocks in the flowchart illustrations and/or block diagrams, can be implemented by computer readable program instructions.

[00173] These computer readable program instructions may be provided to a processor of a general purpose computer, special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer or other programmable data processing apparatus, create means for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks. These computer readable program instructions may also be stored in a computer readable storage medium that can direct a computer, a programmable data processing apparatus, and/or other devices to function in a particular manner, such that the computer readable storage medium having instructions stored therein comprises an article of manufacture including instructions which implement aspects of the function/act specified in the flowchart and/or block diagram block or blocks.

[00174] The computer readable program instructions may also be loaded onto a computer, other programmable data processing apparatus, or other device to cause a series of operational steps to be performed on the computer, other programmable apparatus or other device to produce a computer implemented process, such that the instructions which execute on the computer, other programmable apparatus, or other device implement the functions/acts specified in the flowchart and/or block diagram block or blocks.

[00175] The flowchart and block diagrams in the Figures illustrate the architecture, functionality, and operation of possible implementations of systems, methods, and computer program products according to various embodiments of the present disclosure. In this regard, each block in the flowchart or block diagrams may represent a module, segment, or portion of

instructions, which comprises one or more executable instructions for implementing the specified logical function(s). In some alternative implementations, the functions noted in the block may occur out of the order noted in the Figures. For example, two blocks shown in succession may, in fact, be executed substantially concurrently, or the blocks may sometimes be executed in the reverse order, depending upon the functionality involved. It will also be noted that each block of the block diagrams and/or flowchart illustration, and combinations of blocks in the block diagrams and/or flowchart illustration, can be implemented by special purpose hardware-based systems that perform the specified functions or acts or carry out combinations of special purpose hardware and computer instructions.

We claim:

1. A method comprising:

extracting DNA from a collection of samples from an organism;

preparing a sequence library with duplex adapters, wherein the sequence library is prepared by ligating a duplex adapter having a Unique Molecule Identifier (UMI) to an end of each of a plurality of strands of the extracted DNA and amplifying the extracted DNA with a first PCR;

selecting a subset of the sequence library;

amplifying the subset with a second PCR to increase a number of PCR duplicates;

sequencing a plurality of duplex reads from the amplified subset;

aligning the plurality of duplex reads to a host genome and denoising the plurality of duplex reads based on said alignment;

detecting the presence of a variant in at least one of the plurality of duplex reads;

determining a signature of the variant;

comparing the signature of the variant to a collection of disease-specific variant signatures; and

determining a disease type based on the comparison.

2. The method of claim 1, wherein the UMI has exactly three base pairs.

3. The method of claim 1, wherein the UMI has less than five base pairs.

4. The method of claim 1, wherein the disease type is a cancer type.

5. The method of claim 1, wherein the cancer type comprises bladder cancer.

6. The method of claim 1, wherein the sequence library comprises one of a whole genome library or a whole exome library.

7. The method of claim 6, wherein preparing the sequence library with duplex adapters further comprises collapsing one or more errors on a strand of extracted DNA.

8. The method of claim 1, wherein amplifying the extracted DNA with the first PCR comprises removing sequencing errors based on the presence of two or more molecules with the same UMI.

9. The method of claim 1, wherein sequencing the plurality of duplex reads comprises sequencing on a paired-end system.

10. The method of claim 1, wherein

sequencing the plurality of duplex reads comprises sequencing on a single-end system and, wherein

aligning the plurality of duplex reads to the host genome comprises:

obtaining a plurality of single-end DNA sequencing reads;

separating a top-mapping strand of the DNA from a bottom-mapping strand of DNA;

performing error collapsing on each of the top-mapping strands and the bottom-mapping strands;

reverting the bottom-mapping strands to top-mapping strands by re-grouping based on UMI; and

performing error correcting between the top and bottom strands.

11. The method of claim 1, wherein

sequencing the plurality of duplex reads comprises sequencing on a single-end system and, wherein

aligning sequences to a host genome comprises:

obtaining a plurality of single-end DNA sequencing reads;

creating a synthetic paired-end read; and

performing error correcting on all strands.

12. The method of claim 1, wherein sequencing the plurality of duplex reads comprises sequencing a series of uncorrected reads belonging to a duplex family.

13. The method of claim 12, further comprising processing uncorrected reads belonging to a duplex family to measure a read specific feature.

14. The method of claim 13, further comprising filtering the uncorrected reads based on the measured read specific feature.

15. The method of claim 1, further comprising trimming the sequence of reads from the amplified subset.

16. The method of claim 1, wherein comparing the extracted variant to a collection of cancer-specific variant signatures comprises:

calculating a tumor fraction estimation of a duplex-corrected signature;

plotting the tumor fraction estimation to create a duplex-corrected signature for the extracted variant; and

matching the duplex-corrected signature to a reference signature.

17. The method of claim 16, wherein the signature comprises a relative proportion of a trinucleotide mutation.

18. The method of claim 4, further comprising:

correcting for library and sequencing artifacts by a panel of cancer-free controls sequenced on the same system; and

estimating a tumor fraction.

19. The method of claim 1, further comprising integrating a genome-wide mutation from the sequencing reads as a weighted sum of single-base substitution (SBS) reference mutational signatures.

20. The method of claim 19, wherein integrating the genome-wide mutation comprises:

deconvolving SBS mutational signatures from plasma DNA mixtures using a non-negative maximum likelihood model;

estimating a tumor fraction by taking a weight of a tumor-associated SBS signature and normalizing by a total number of mutations and depth of sequencing; and

calculating a signature score to determine that the cancer-associated SCS explains an observed mutation profile.

21. The method of claim 1, further comprising discarding a variant with an allele frequency greater than 30%.

22. The method of claim 21, further comprising:

aggregating reads having a variant with allele frequency less than 30%;

calculating a frequency of variants in the aggregated reads trinucleotide context; and

comparing the calculated trinucleotide variant frequency with a number of reference frequencies for different biological processes.

23. The method of claim 1, wherein determining a disease status comprises:

randomly changing a trinucleotide frequency of a reference signature from the collection;

performing a non-negative maximum likelihood fit between the randomly-permutated trinucleotide frequency and a frequency of the signature; and

scoring the fit below a disease-negative threshold.

24. The method of claim 1, wherein the DNA is genomic DNA.

25. The method of claim 1, wherein the DNA is cell-free DNA (cfDNA).

26. The method of claim 1, wherein detecting the presence of the variant comprises:

providing the plurality of duplex reads to a pretrained machine learning model; and

receiving therefrom an indication of a base variant irrespective of comparative sequence length.

27. The method of claim 26, wherein the pretrained machine learning model comprises an artificial neural network.

28. The method of claim 27, wherein the artificial neural network is one of a feedforward neural network, a radial basis function network, a self-organizing map, learning vector quantization, a recurrent neural network, a Hopfield network, a Boltzmann machine, an echo state network, long short term memory, a bi-directional recurrent neural network, a hierarchical recurrent neural network, a stochastic neural network, a modular neural network, an associative neural network, a deep neural network, a deep belief network, a convolutional

neural networks, a convolutional deep belief network, a large memory storage and retrieval neural network, a deep Boltzmann machine, a deep stacking network, a tensor deep stacking network, a spike and slab restricted Boltzmann machine, a compound hierarchical-deep model, a deep coding network, a multilayer kernel machine, or a deep Q-network.

29. The method of claim 26, wherein the pretrained machine learning model comprises a trained classifier.

30. The method of claim 29, wherein the trained classifier is a random decision forest.

31. The method of claim 1, wherein the collection of samples comprises a collection of plasma samples.

32. The method of claim 1, wherein the sequence library comprises a whole genome sequence library.

33. A computer program product for reducing sequencing error rates, the computer program product comprising a computer readable storage medium having program instructions embodied therewith, the program instructions executable by a processor to cause the process to perform a method according to any one of claims 1-32.

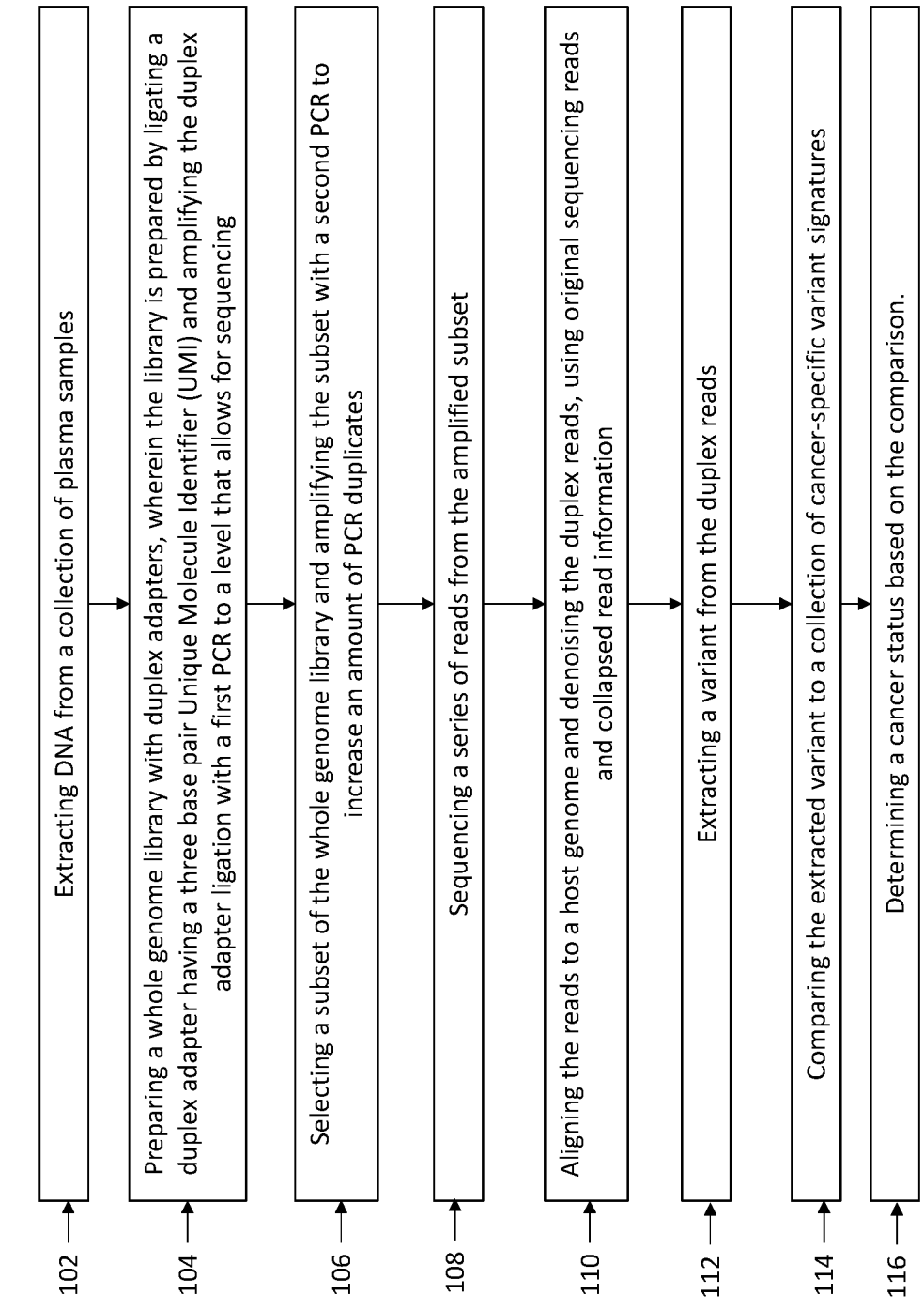

100
102
Extracting DNA from a collection of plasma samples
104
Preparing a whole genome library with duplex adapters, wherein the library is prepared by ligating a duplex adapter having a three base pair Unique Molecule Identifier (UMI) and amplifying the duplex adapter ligation with a first PCR to a level that allows for sequencing
106
Selecting a subset of the whole genome library and amplifying the subset with a second PCR to increase an amount of PCR duplicates
108
Sequencing a series of reads from the amplified subset
110
Aligning the reads to a host genome and denoising the duplex reads, using original sequencing reads and collapsed read information
112
Extracting a variant from the duplex reads
114
Comparing the extracted variant to a collection of cancer-specific variant signatures
116
Determining a cancer status based on the comparison.

FIG. 1

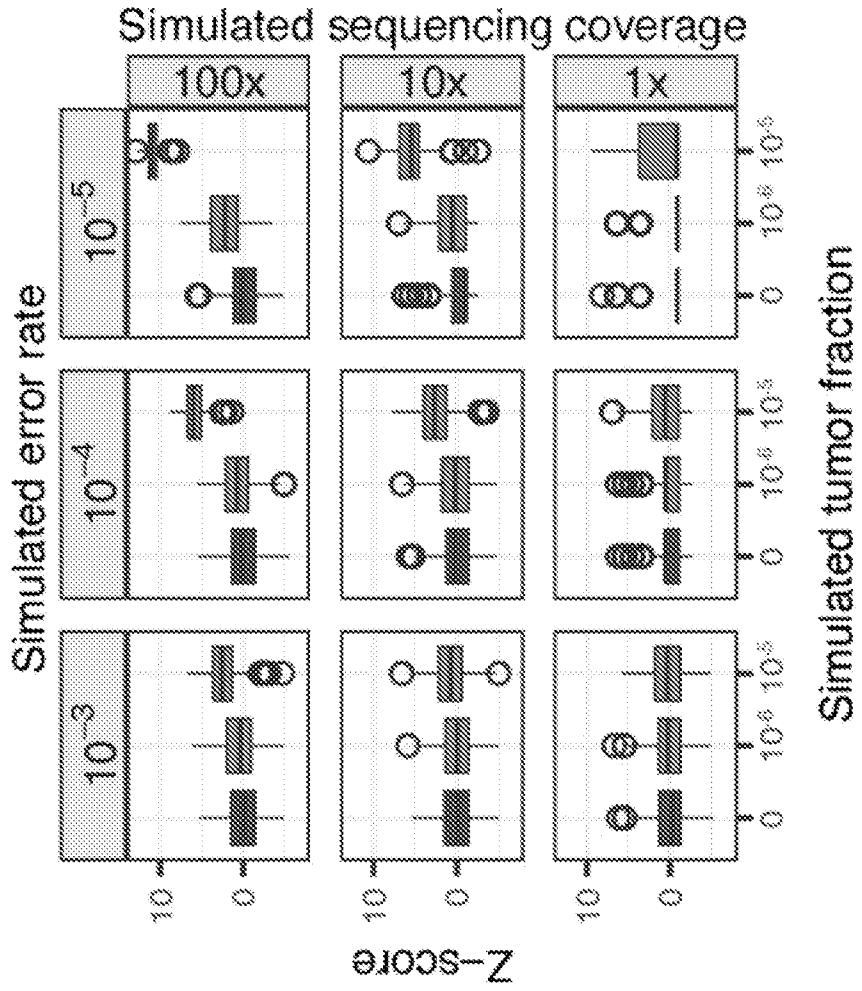

Simulated sequencing coverage
100x
10x
1x
Simulated error rate
10^{-5}
10^{-4}
10^{-3}
Simulated tumor fraction
0
10^{-6}
10^{-5}
Z-score
10
0

FIG. 2A

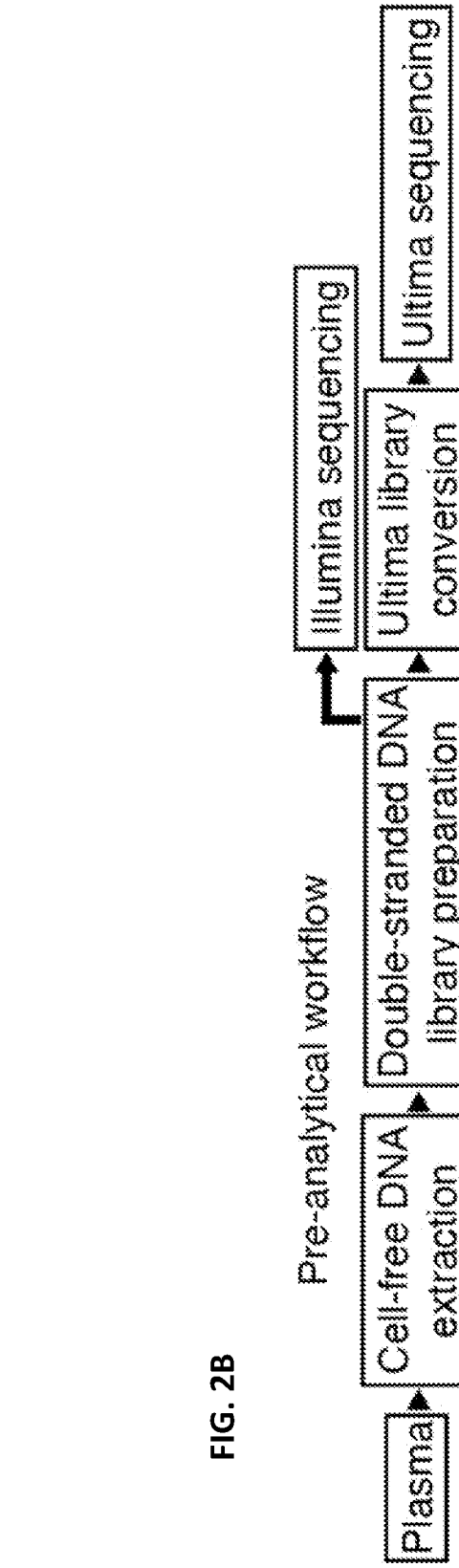

Plasma
Cell-free DNA extraction
Pre-analytical workflow
Double-stranded DNA library preparation
Illumina sequencing
Ultima library conversion
Ultima sequencing

FIG. 2B

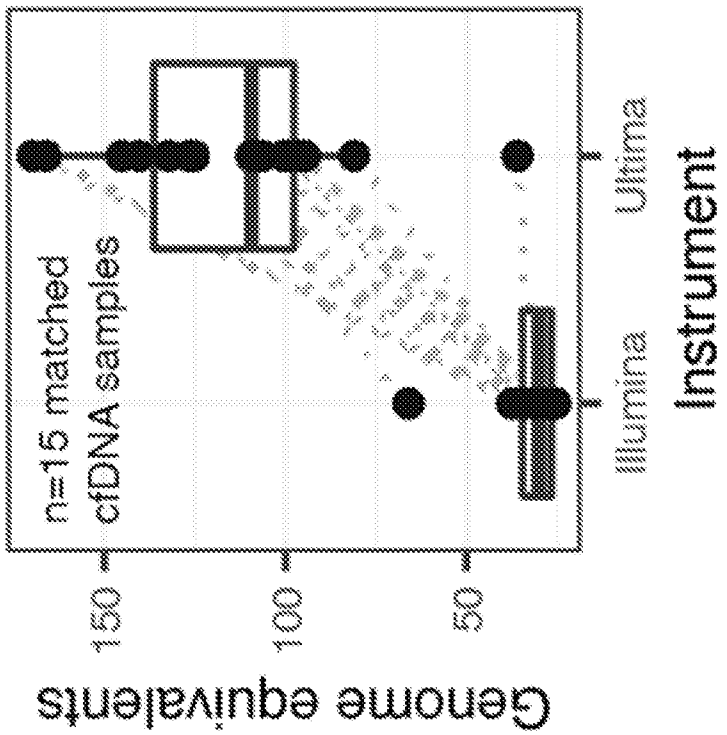

Genome equivalents
150
100
50
n=15 matched
cfDNA samples
Illumina
Ultima
Instrument

FIG. 2C

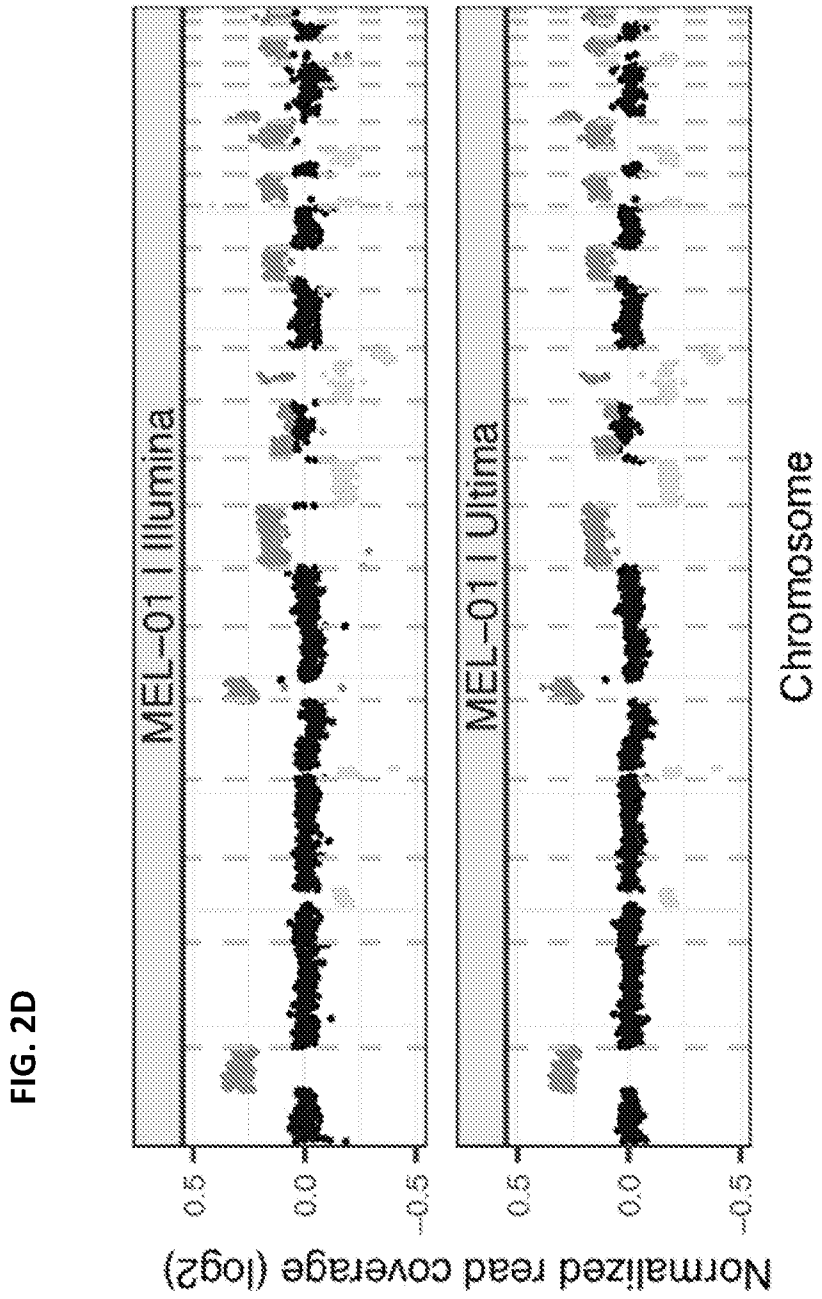

MEL-01 | Illumina
MEL-01 | Ultima
0.5
0.0
−0.5
Normalized read coverage (log2)
Chromosome

FIG. 2D

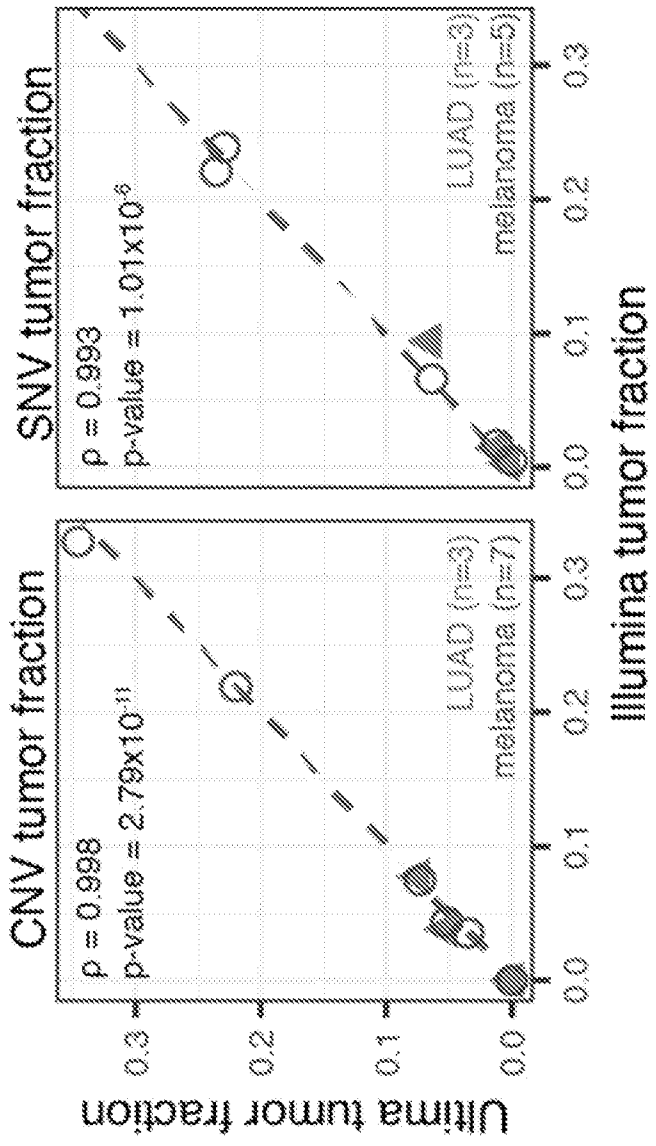
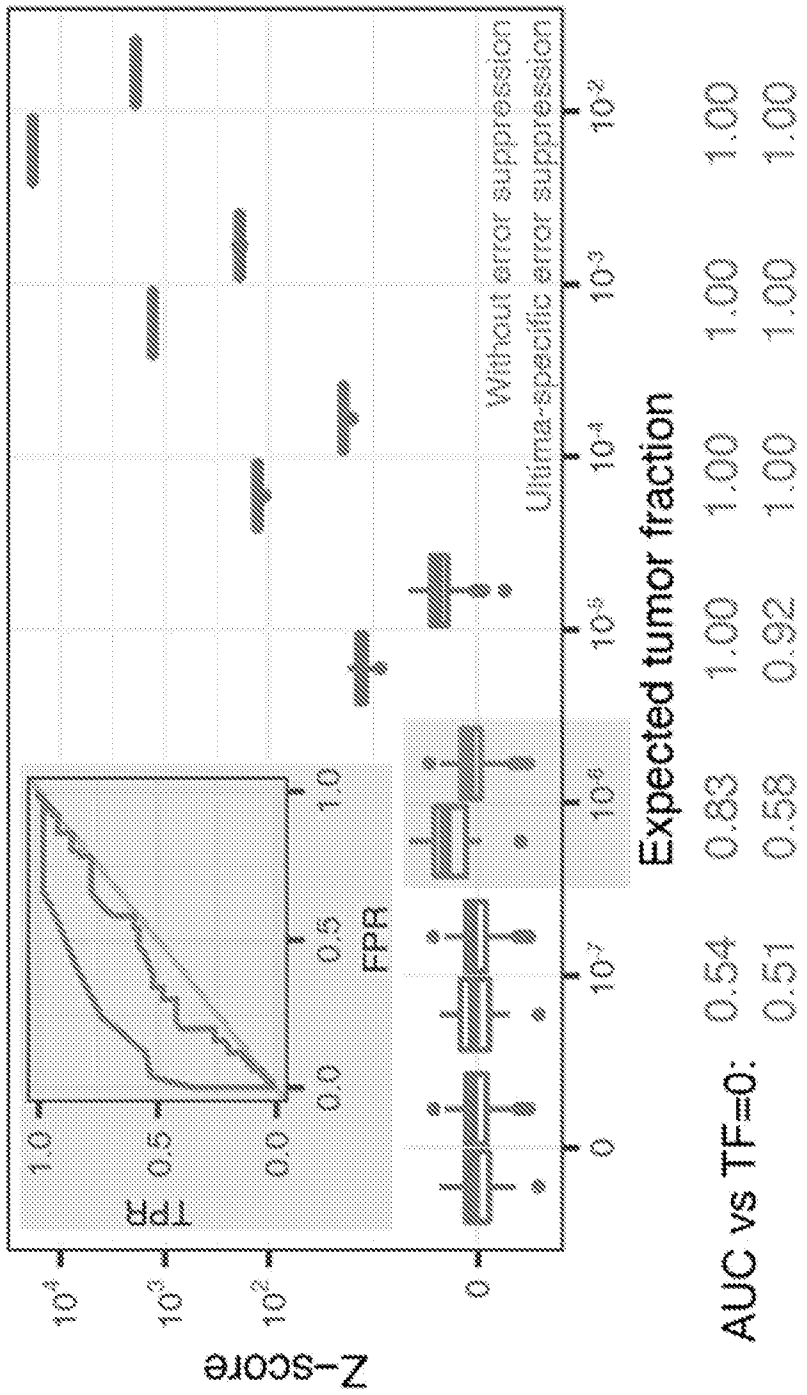

CNV tumor fraction
SNV tumor fraction
ρ = 0.998
p-value = 2.79x10⁻¹¹
ρ = 0.993
p-value = 1.01x10⁻⁶
LUAD (n=3)
melanoma (n=7)
LUAD (n=3)
melanoma (n=5)
Ultima tumor fraction
Illumina tumor fraction
0.0
0.1
0.2
0.3

FIG. 2E

FIG. 2F


Z-score
Expected tumor fraction
Without error suppression
Ultima-specific error suppression
TPR
FPR
AUC vs TF=0:
0.54 0.83 1.00 1.00 1.00 1.00
0.51 0.58 0.92 1.00 1.00 1.00

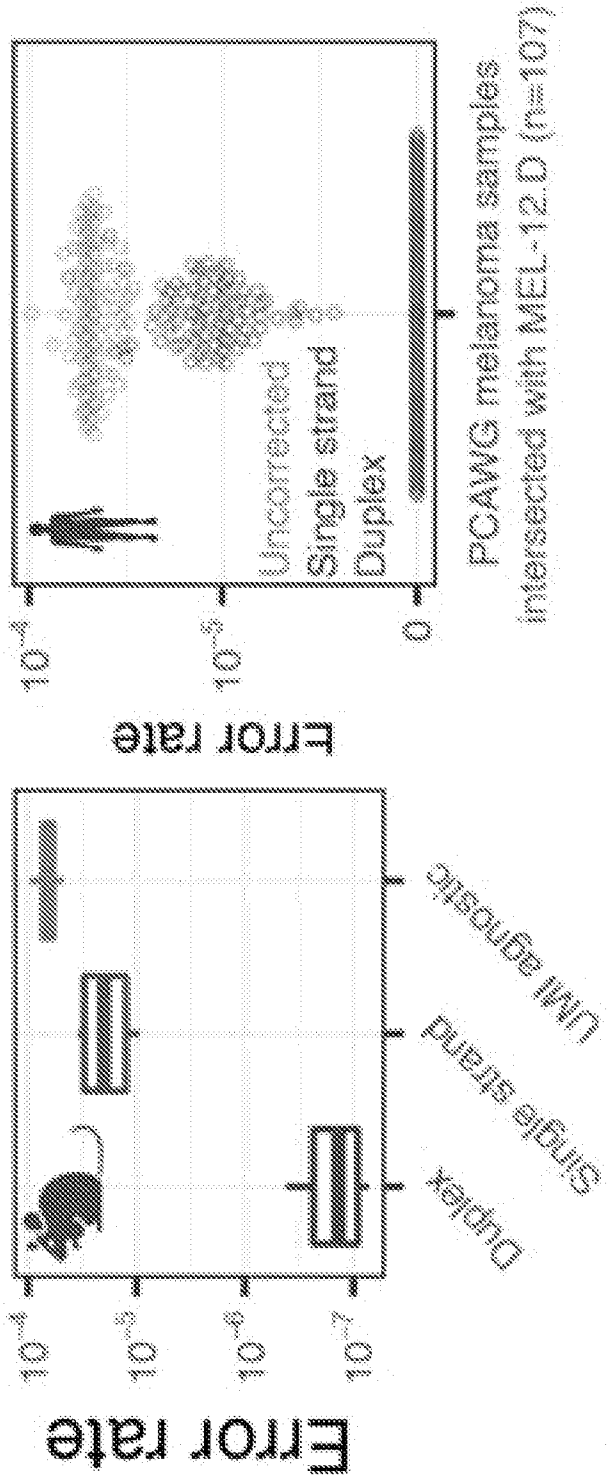
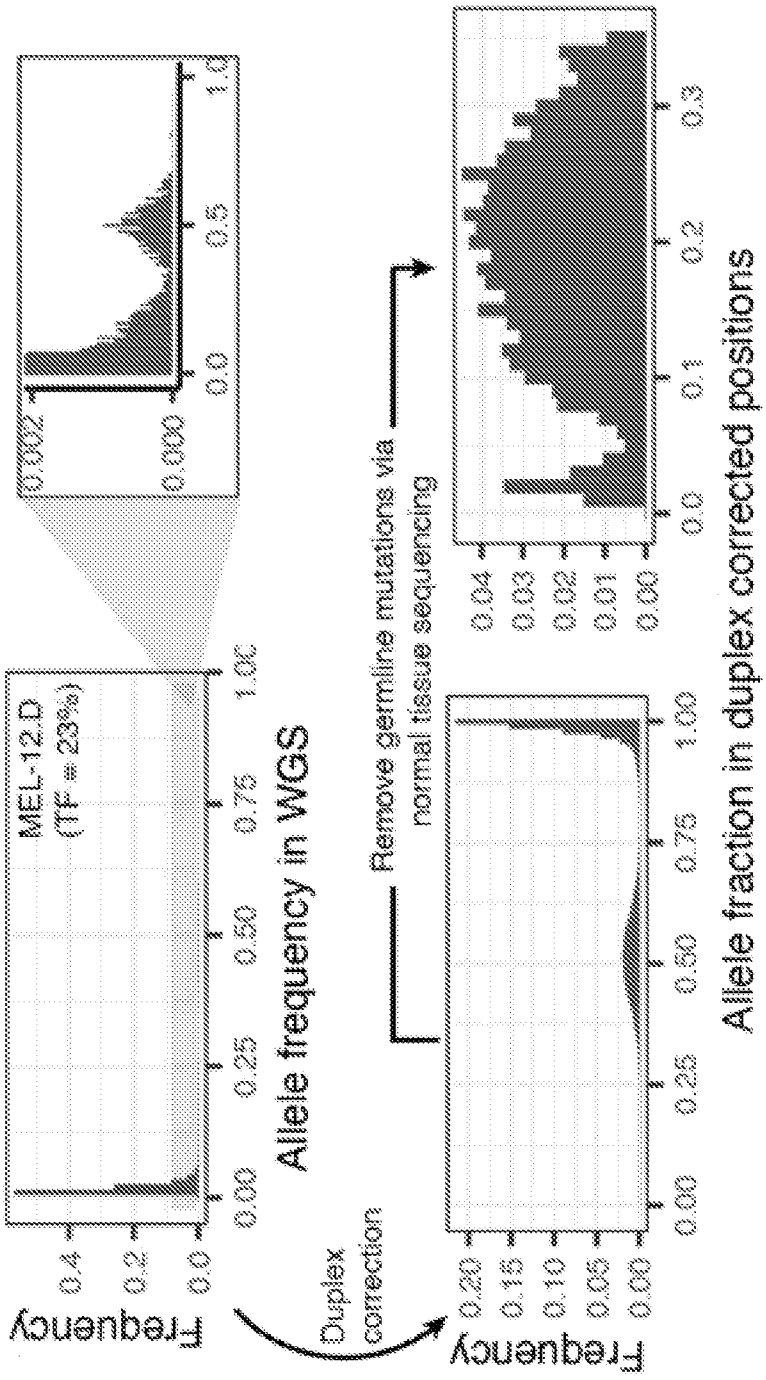

Error rate
10^{-4}
10^{-5}
10^{-6}
10^{-7}
Duplex
Single strand
UMI agnostic
Error rate
10^{-4}
10^{-5}
0
Uncorrected
Single strand
Duplex
PCAWG melanoma samples
intersected with MEL-12.0 (n=107)

FIG. 3A

FIG. 3B


MEL-12.D
(TF = 23%)
Frequency
Allele frequency in WGS
Duplex correction
Remove germline mutations via normal tissue sequencing
Allele fraction in duplex corrected positions

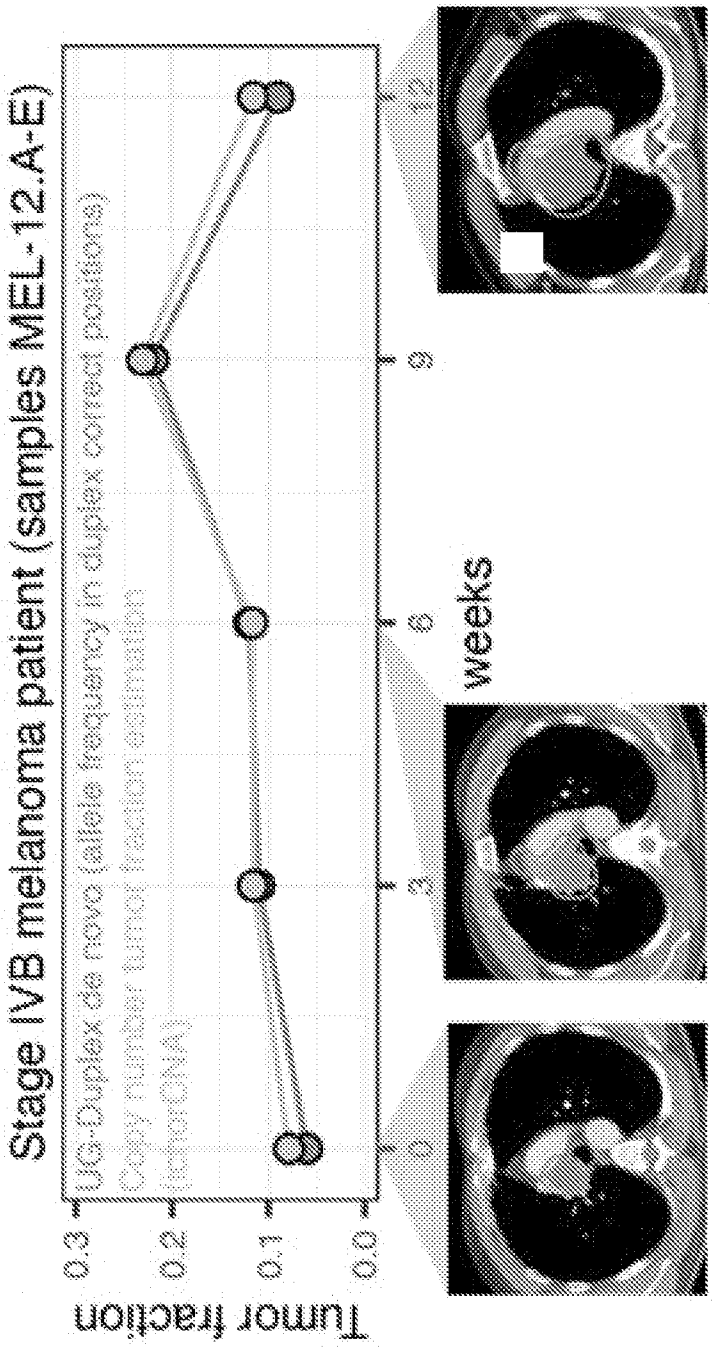

Stage IVB melanoma patient (samples MEL-12.A-E)
UG-Duplex de novo (allele frequency in duplex corrected positions)
Copy number tumor fraction estimation
ichorCNA
Tumor fraction
0.3
0.2
0.1
0.0
0
3
6
9
12
weeks

FIG. 3C

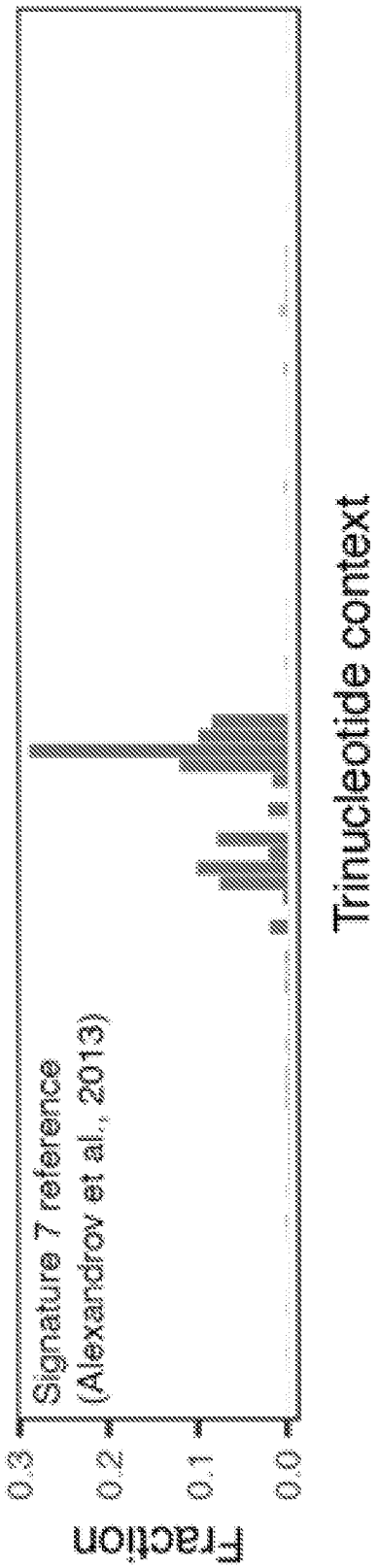

Fraction
0.3
0.2
0.1
0.0
Signature 7 reference
(Alexandrov et al., 2013)
Trinucleotide context

FIG. 3D

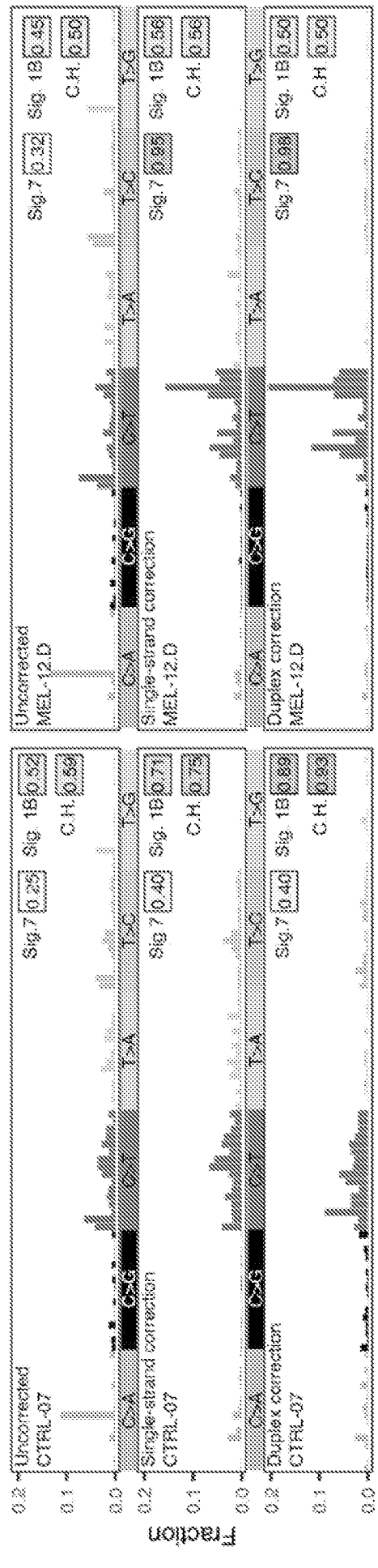

Uncorrected
CTRL-07
Single-strand correction
CTRL-07
Duplex correction
CTRL-07
Uncorrected
MEL-12.D
Single-strand correction
MEL-12.D
Duplex correction
MEL-12.D
Sig.7
Sig. 1B
C.H.
C>A
C>G
T>A
T>C
T>G
Fraction
Trinucleotide context
Cosine similarity

FIG. 3E

FIG. 3F

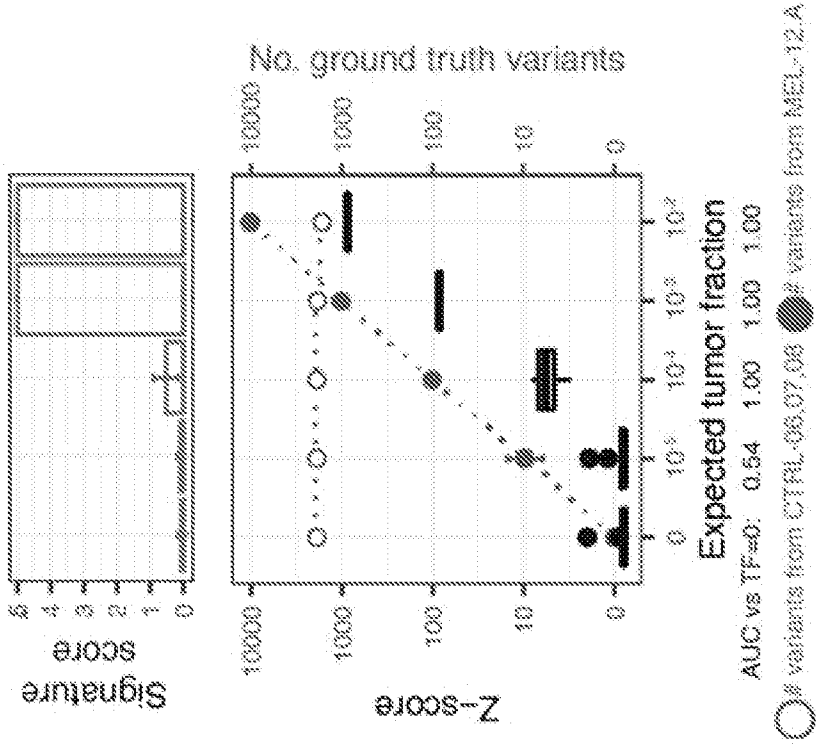

No. ground truth variants
10000
1000
100
10
0
Z-score
Signature score
Expected tumor fraction
AUC vs TF=0: 0.54 1.00 1.00 1.00
variants from CTRL-06.07.08
variants from MEL-12.A

FIG. 3G

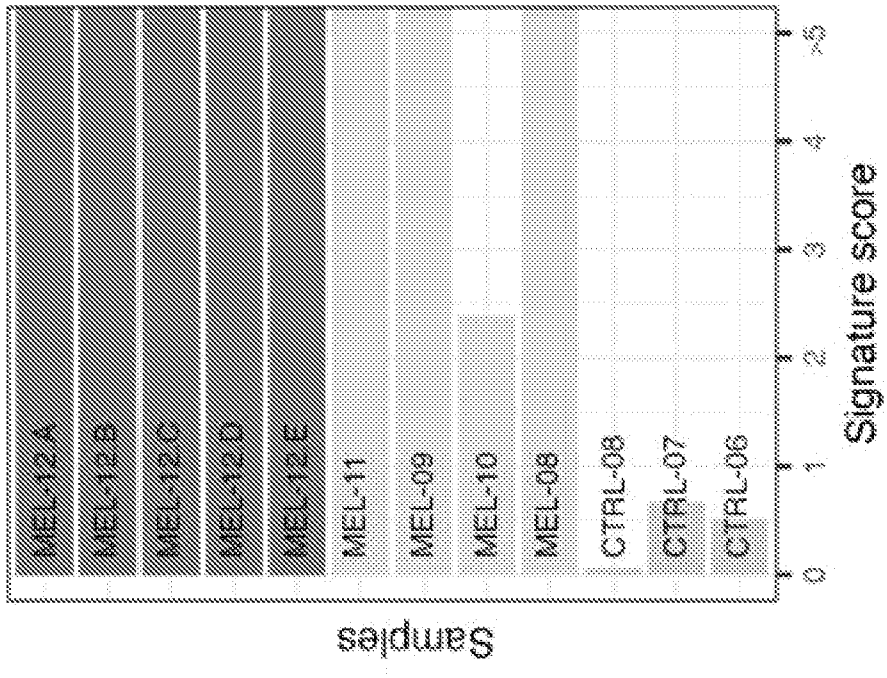
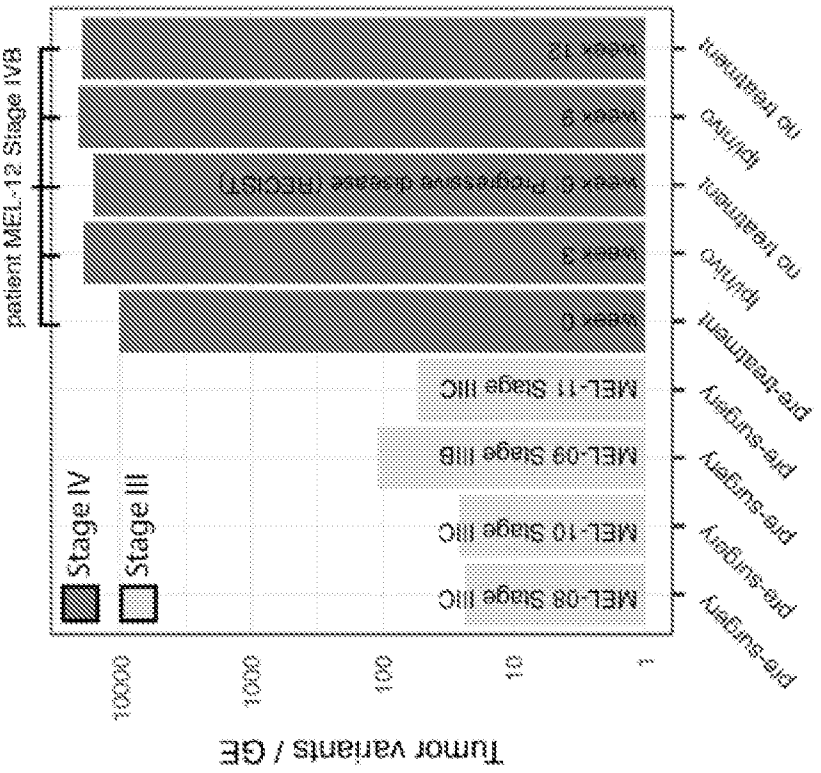

MEL-12A
MEL-12B
MEL-12C
MEL-12D
MEL-12E
MEL-11
MEL-09
MEL-10
MEL-08
CTRL-08
CTRL-07
CTRL-06
Samples
0
1
2
3
4
×5
Signature score

FIG. 3H


Tumor variants / GE
10000
1000
100
10
1
Stage IV
Stage III
patient MEL-12 Stage IVB
MEL-08 Stage IIIC
MEL-10 Stage IIIC
MEL-09 Stage IIIB
MEL-11 Stage IIIC
pre-treatment
pre-surgery
pre-surgery
pre-surgery
pre-surgery
no treatment
on treatment
no treatment
on treatment

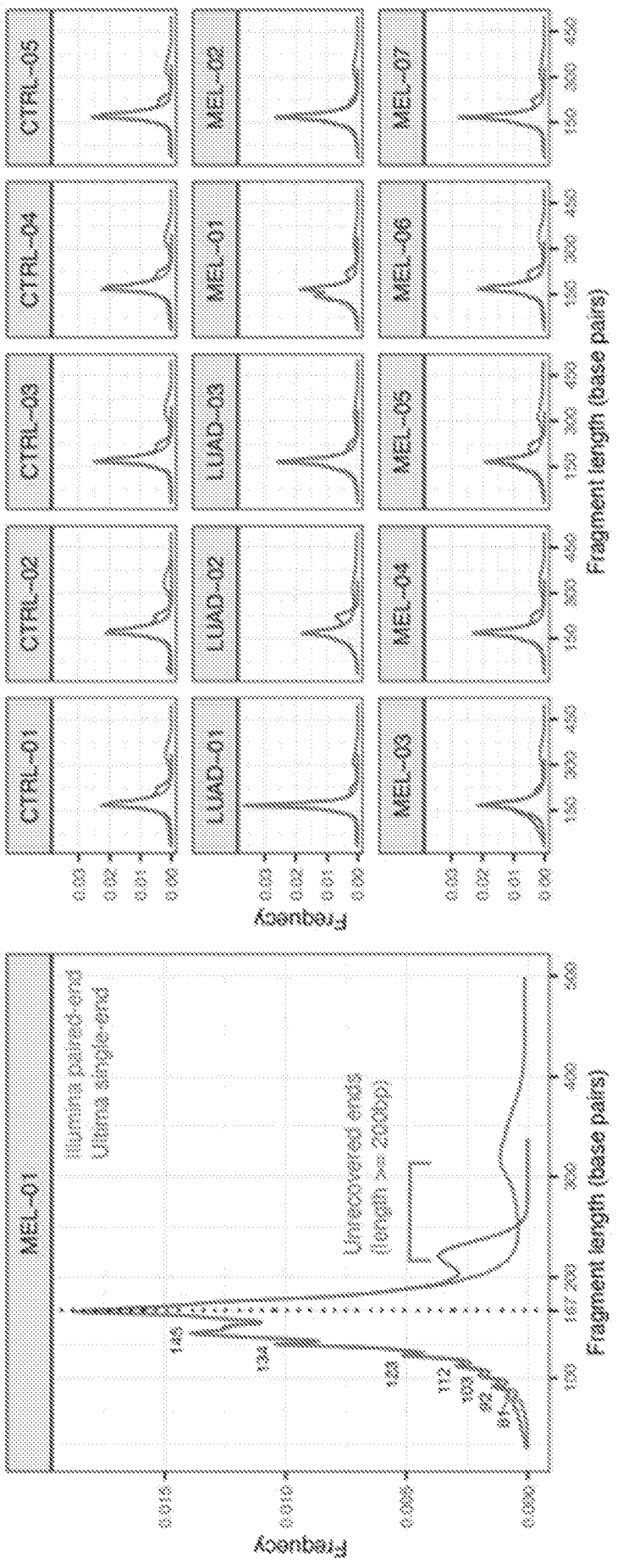

MEL–01
Illumina paired-end
Ultima single-end
Unrecovered ends
(length >= 2000bp)
146
124
159
112
103
82
81
Frequency
Fragment length (base pairs)
CTRL–01
CTRL–02
CTRL–03
CTRL–04
CTRL–05
LUAD–01
LUAD–02
LUAD–03
MEL–01
MEL–02
MEL–03
MEL–04
MEL–05
MEL–06
MEL–07

FIG. 4

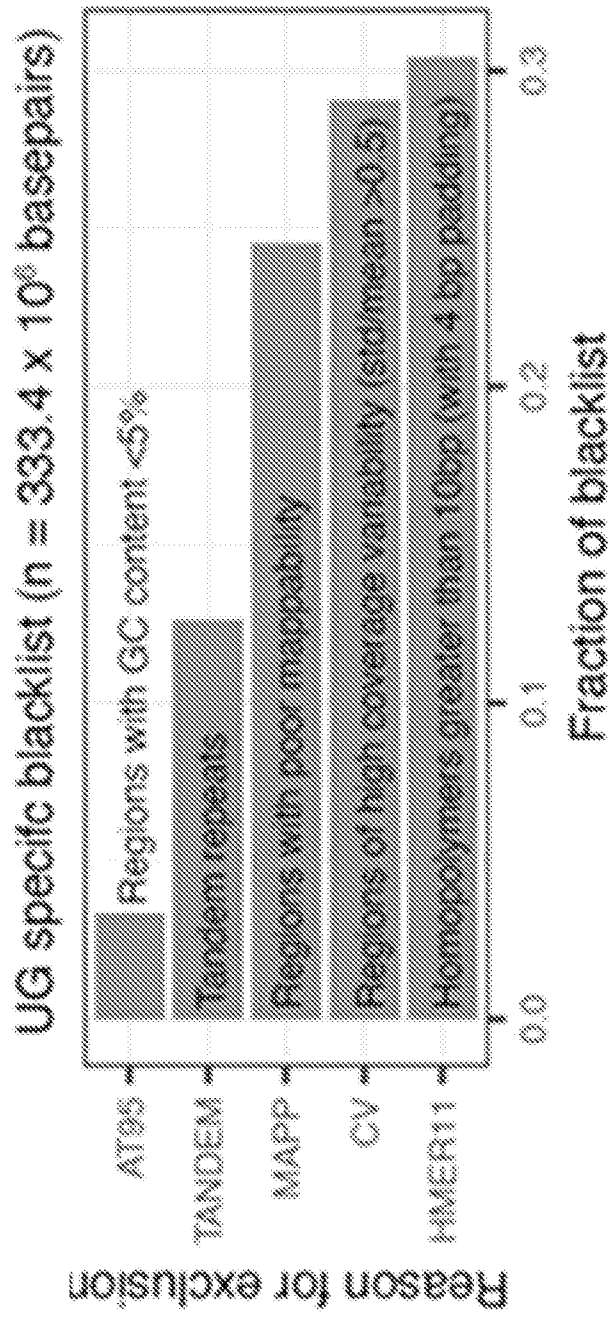

UG specific blacklist (n = 333.4 x 10^6 basepairs)
Reason for exclusion
AT05
TANDEM
MAPP
CV
HMER11
Regions with GC content <5%
Tandem repeats
Regions with poor mappability
Regions of high coverage variability (stdev over 0.5)
Homopolymers greater than 10bp (with 4 bp padding)
0.0
0.1
0.2
0.3
Fraction of blacklist

FIG. 5A

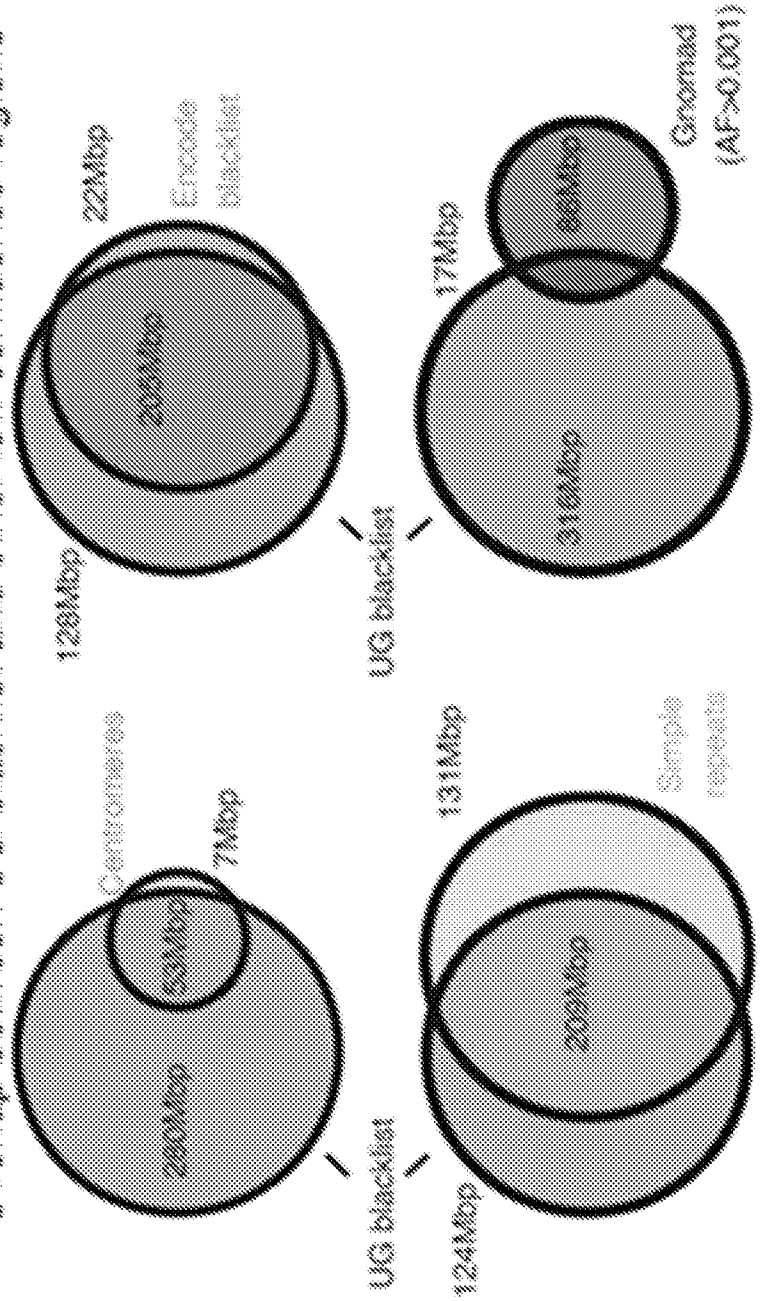

Overlap between UG blacklist and other low confidence regions
UG blacklist
128Mbp
22Mbp
22Mbp
Encode blacklist
UG blacklist
124Mbp
131Mbp
20Mbp
Simple repeats
UG blacklist
316Mbp
17Mbp
Gnomad (AF>0.001)

FIG. 5B

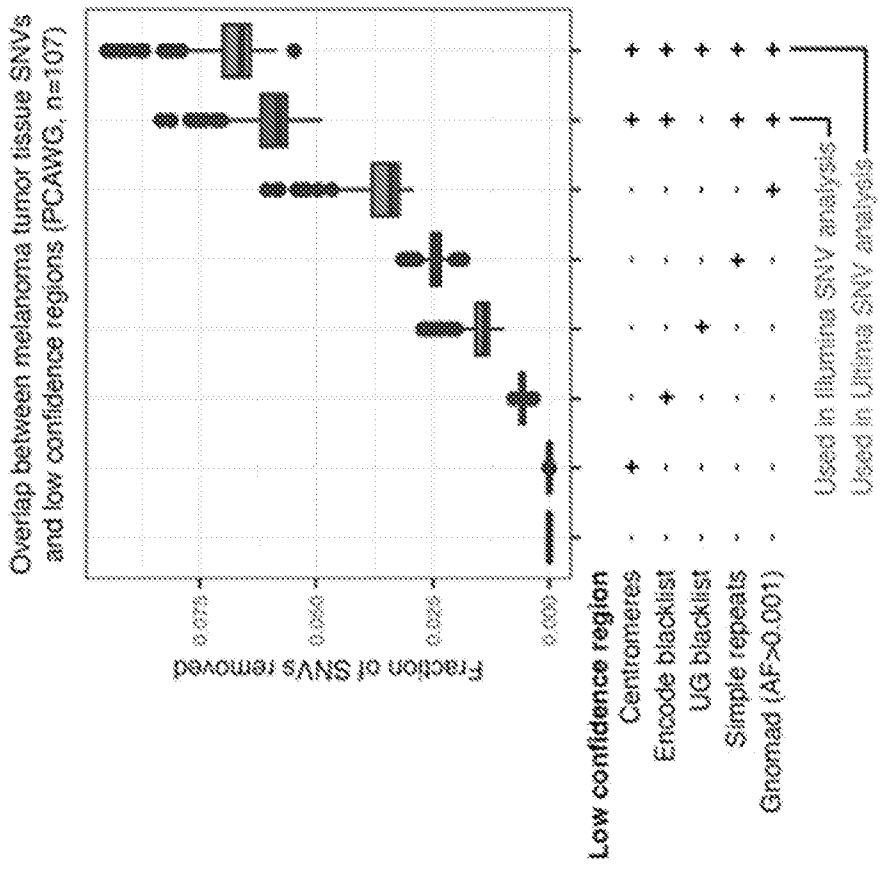

Overlap between melanoma tumor tissue SNVs and low confidence regions (PCAWG, n=107)
Fraction of SNVs removed
0.015
0.010
0.005
0.000
Low confidence region
Centromeres
Encode blacklist
UG blacklist
Simple repeats
Gnomad (AF>0.001)
Used in Illumina SNV analysis
Used in Ultima SNV analysis

FIG. 5C

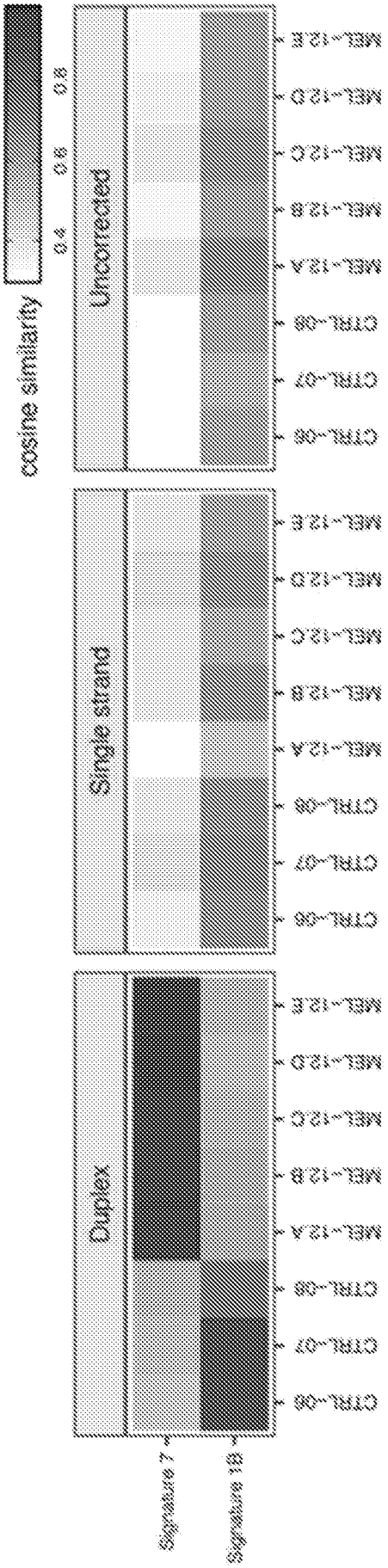

cosine similarity
0.4
0.6
0.8
Uncorrected
Single strand
Duplex
CTRL-06
CTRL-07
CTRL-08
MEL-12.A
MEL-12.B
MEL-12.C
MEL-12.D
MEL-12.E
Signature 7
Signature 18

FIG. 6A

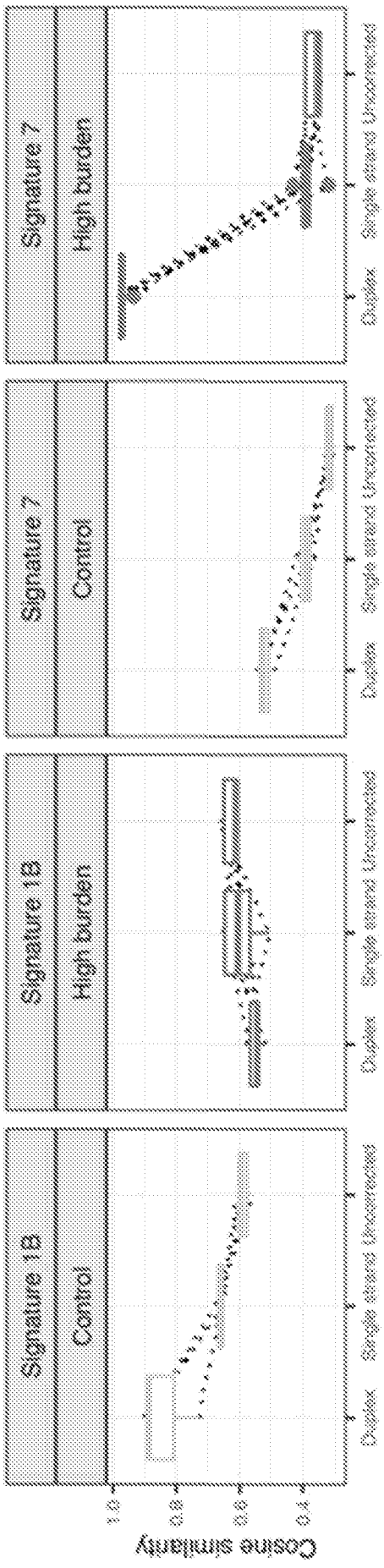

Signature 18
Control
Signature 18
High burden
Signature 7
Control
Signature 7
High burden
Cosine similarity
1.0
0.8
0.6
0.4
Duplex
Single strand Uncorrected
Duplex
Single strand Uncorrected
Duplex
Single strand Uncorrected
Duplex
Single strand Uncorrected

FIG. 6B

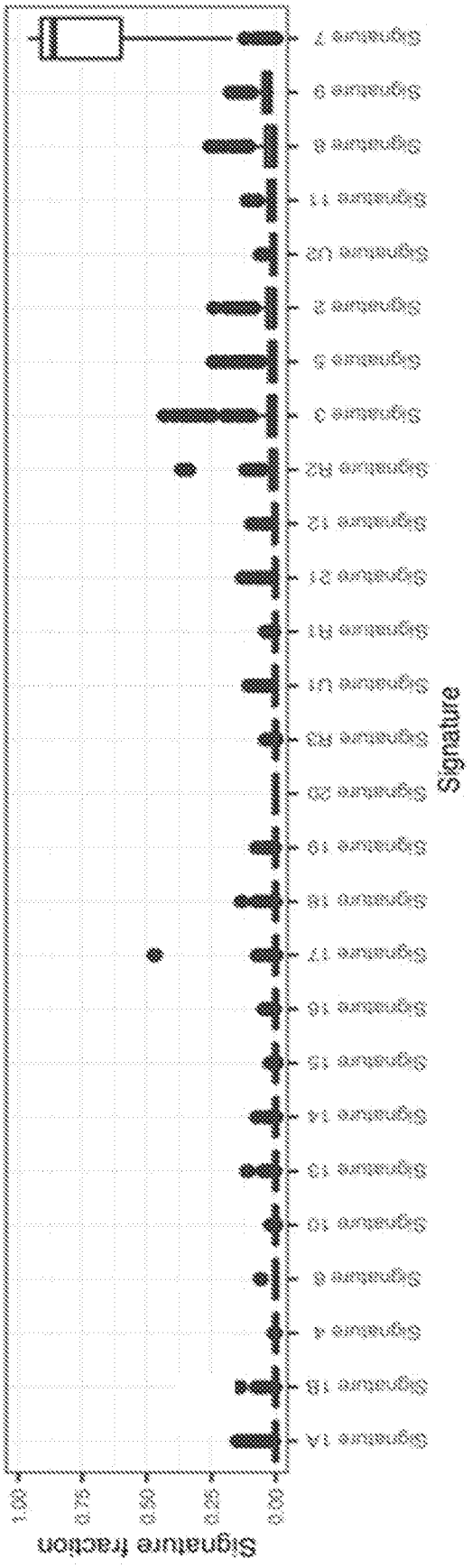

Signature fraction
0.00
0.25
0.50
0.75
1.00
Signature
Signature 1A
Signature 1B
Signature 4
Signature 6
Signature 10
Signature 13
Signature 14
Signature 15
Signature 16
Signature 17
Signature 18
Signature 19
Signature 20
Signature R3
Signature U1
Signature R1
Signature 21
Signature 12
Signature R2
Signature 3
Signature 5
Signature 2
Signature U2
Signature 11
Signature 8
Signature 9
Signature 7

FIG. 7A

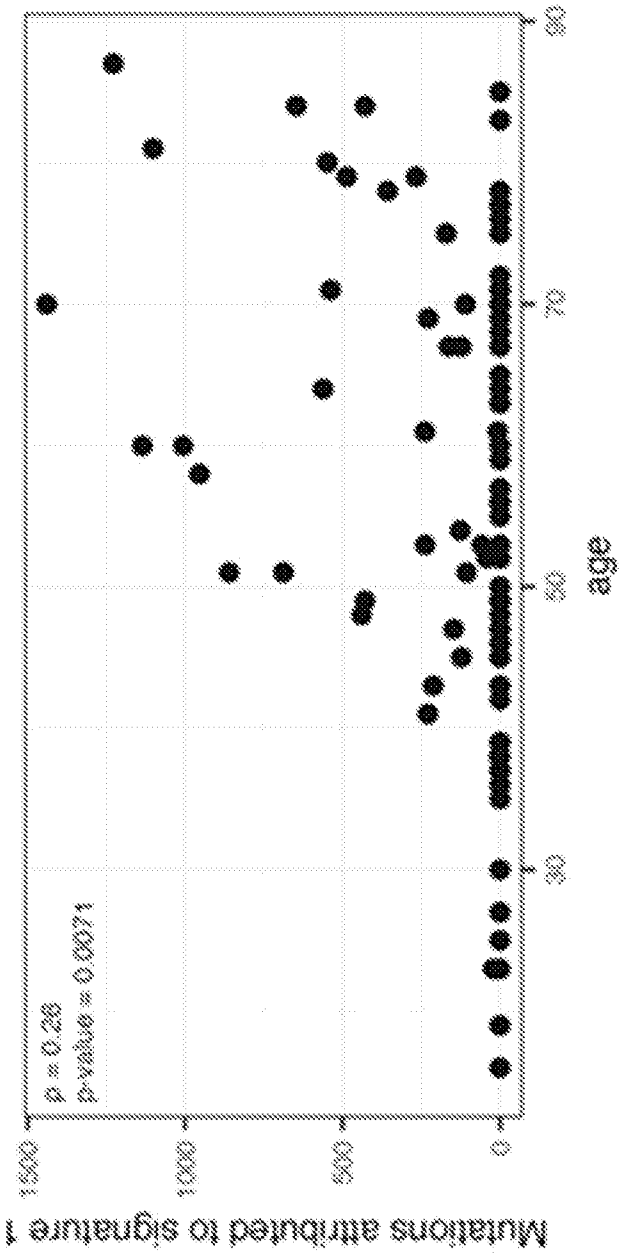

Mutations attributed to signature 1
1500
1000
500
0
ρ = 0.28
p-value = 0.0071
30
50
70
90
age

FIG. 7B

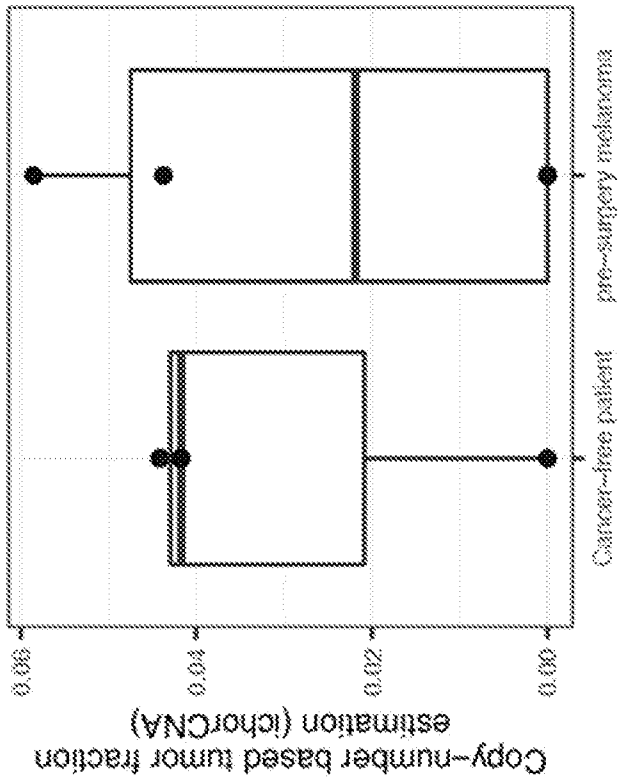

Copy–number based tumor fraction
estimation (ichorCNA)
0.00
0.01
0.02
0.03
Cancer–free patient
pre–surgery melanoma

FIG. 8

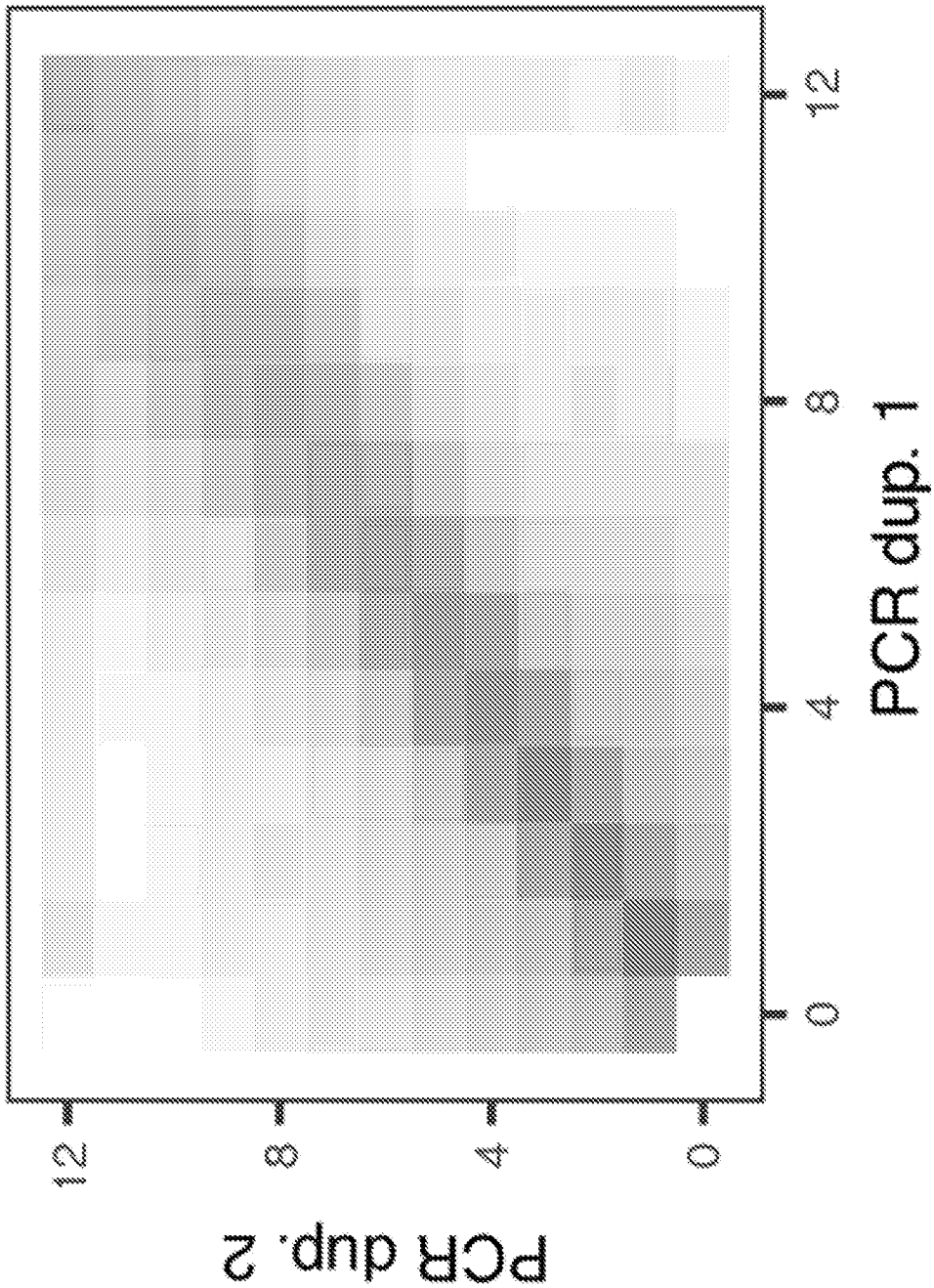


FIG. 9A

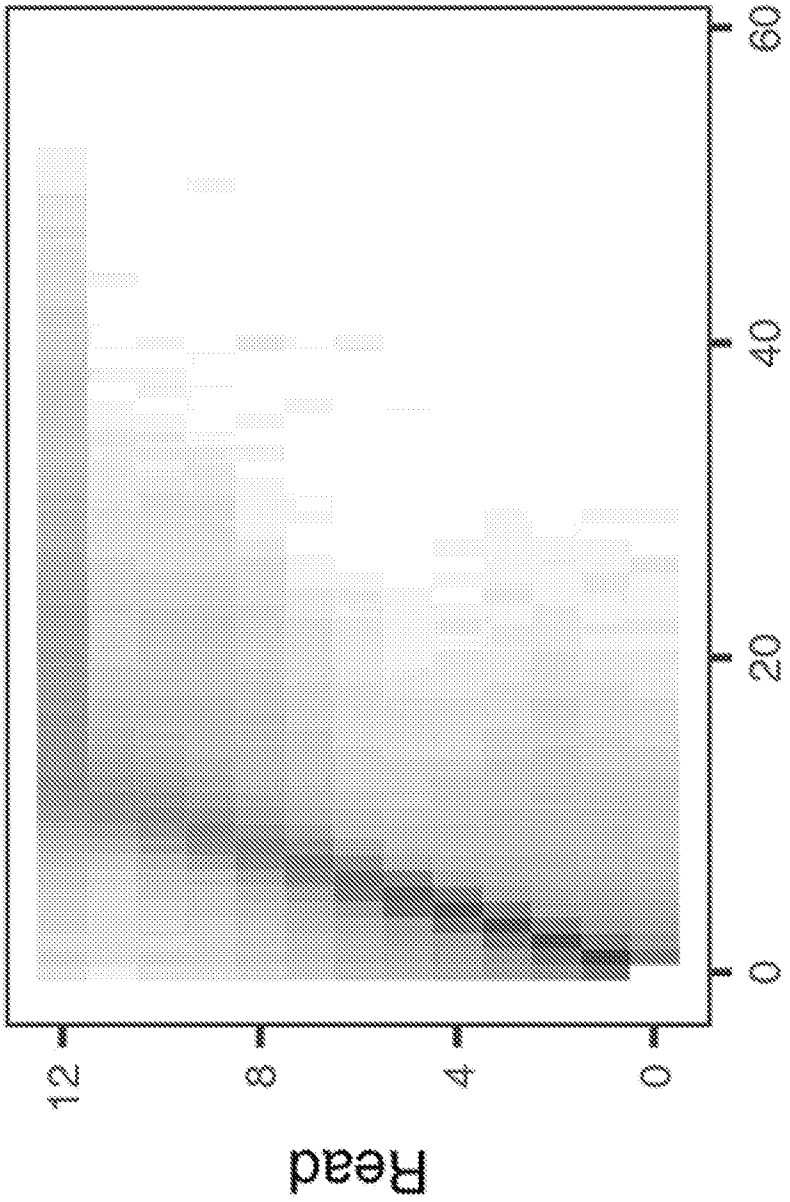


FIG. 9B

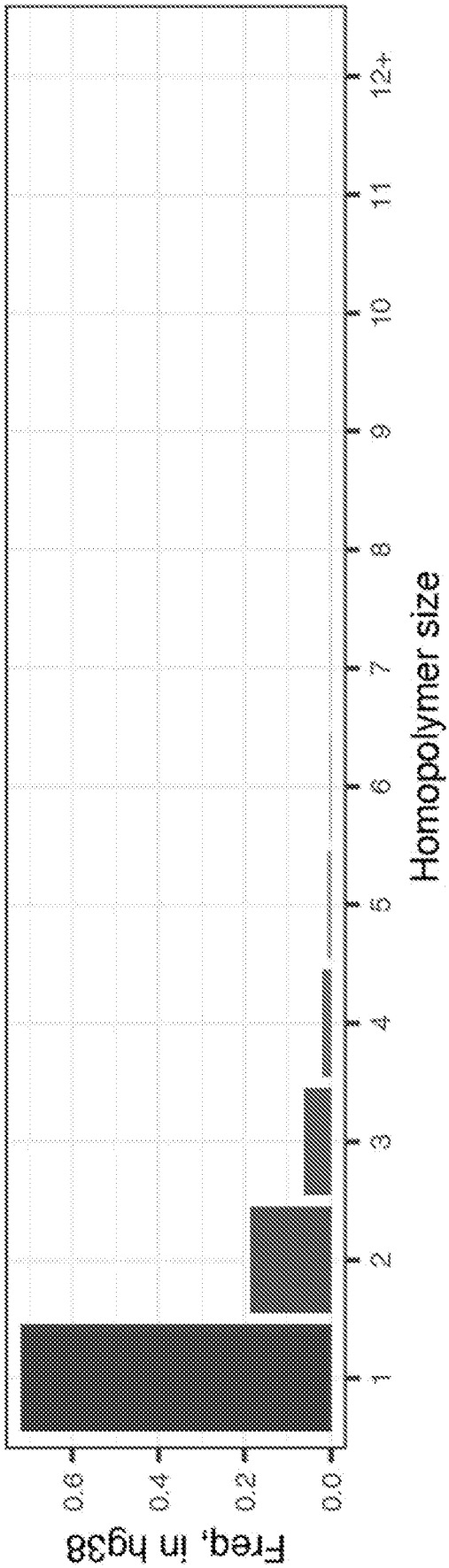


FIG. 9C

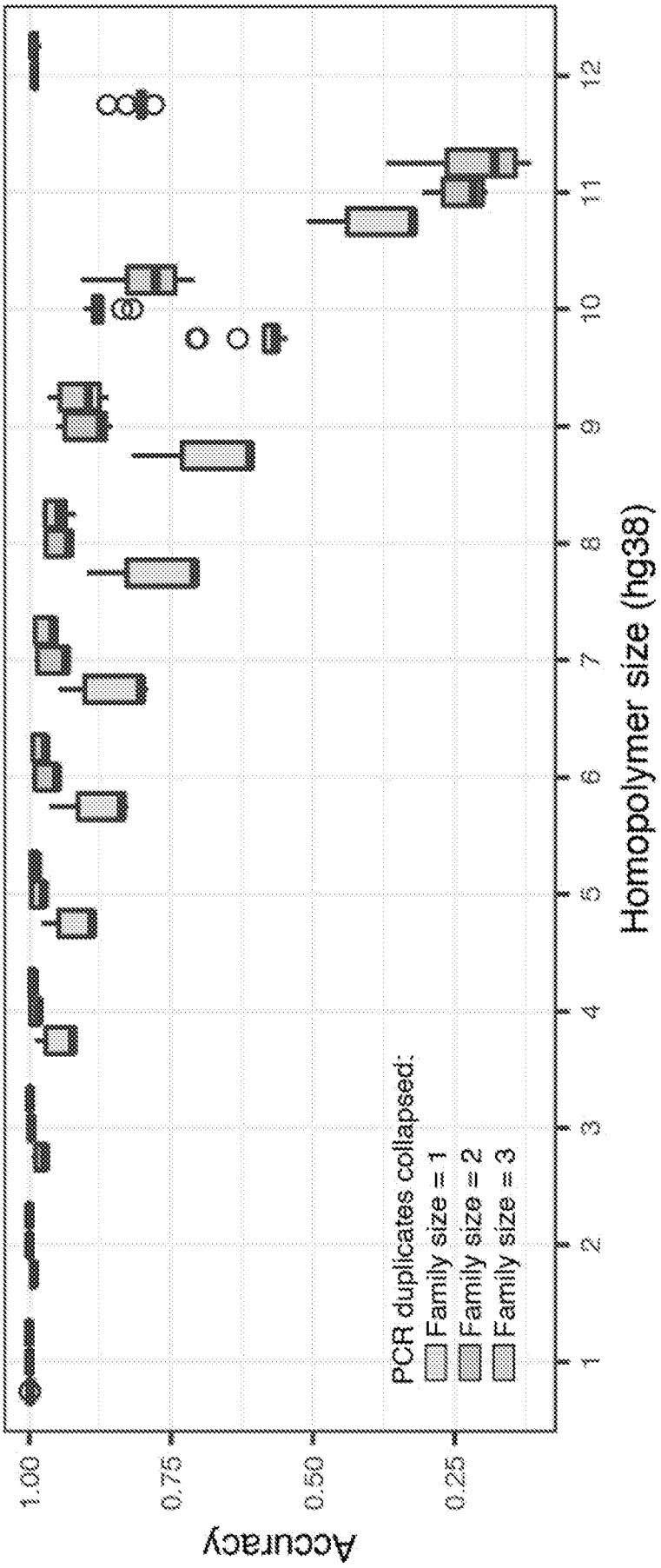

Accuracy
1.00
0.75
0.50
0.25
1
2
3
4
5
6
7
8
9
10
11
12
Homopolymer size (hg38)
PCR duplicates collapsed:
Family size = 1
Family size = 2
Family size = 3

FIG. 9D

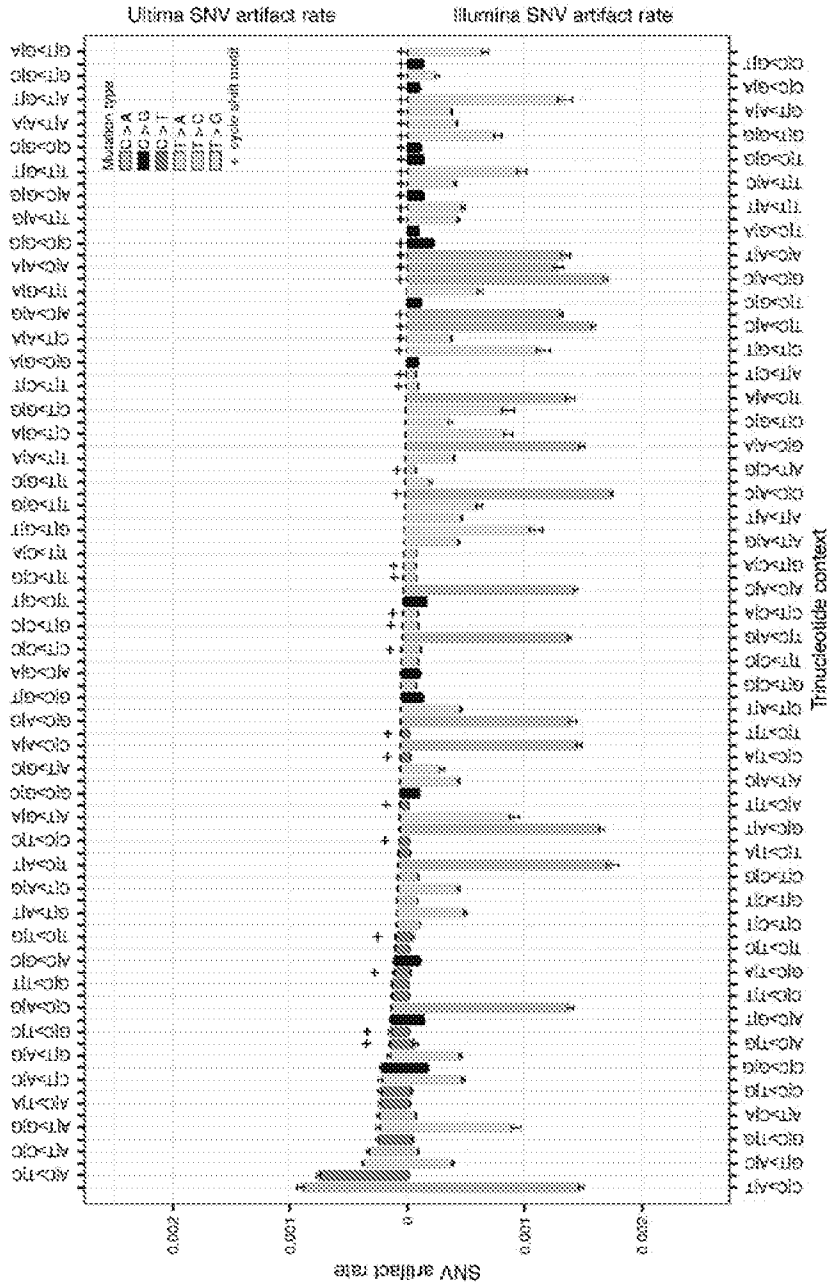

Ultima SNV artifact rate
Illumina SNV artifact rate
Mutation type
C > A
C > G
C > T
T > A
T > C
T > G
+ cycle shift model
SNV artifact rate
0.002
0.001
0
0.001
0.002
Trinucleotide context

FIG. 10

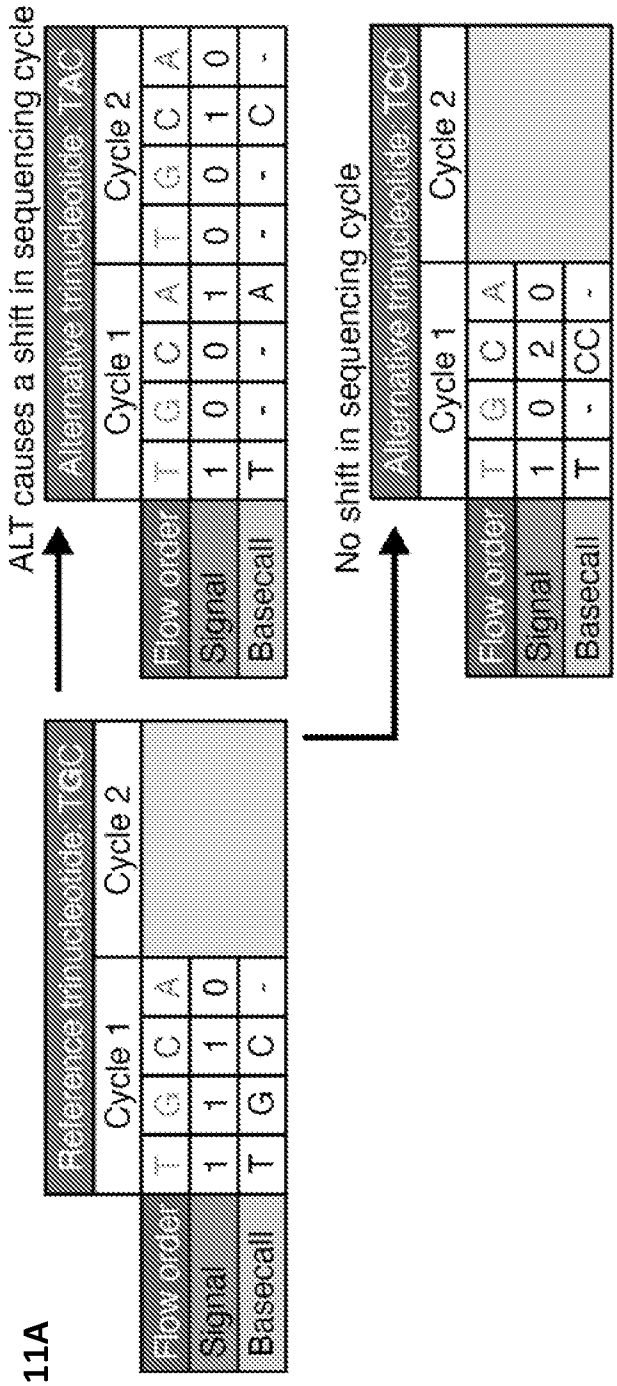

Reference nucleotide TCC
Cycle 1
Cycle 2
Flow order
T G C A
Signal
1 1 1 0
Basecall
T G C -
ALT causes a shift in sequencing cycle
Alternative nucleotide TAC
Cycle 1
Cycle 2
Flow order
T G C A T G C A
Signal
1 0 0 1 0 0 1 0
Basecall
T - - A - - C -
No shift in sequencing cycle
Alternative nucleotide TCC
Cycle 1
Cycle 2
Flow order
T G C A
Signal
1 0 2 0
Basecall
T - CC -

FIG. 11A

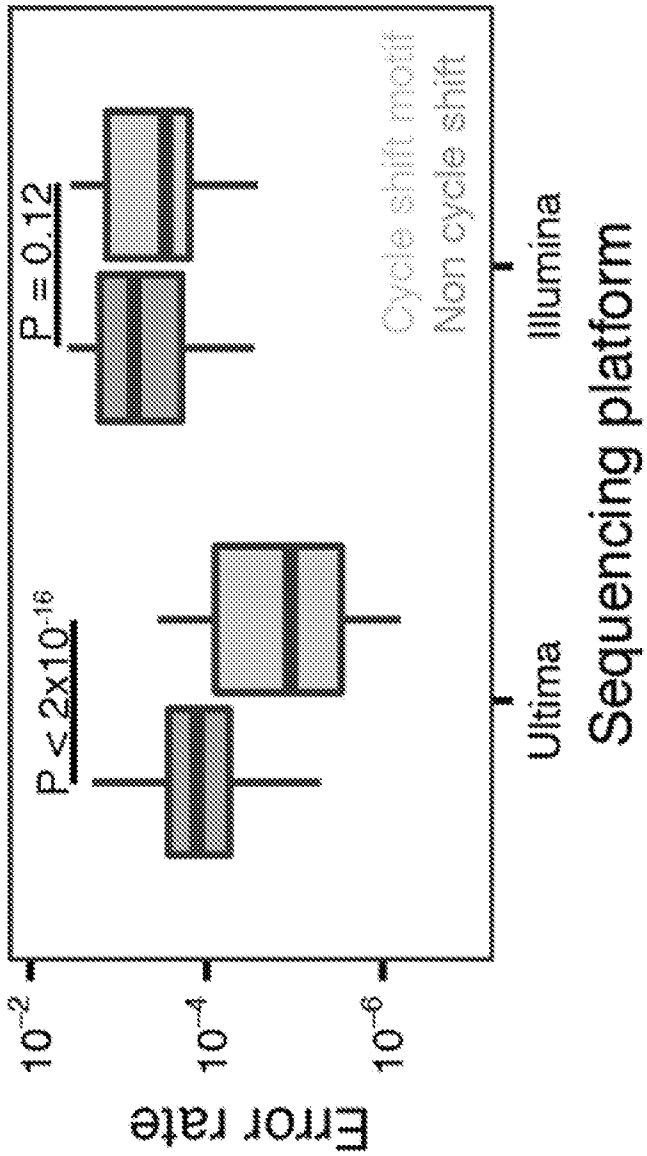

Error rate
10^{-2}
10^{-4}
10^{-6}
P < 2x10^{-16}
P = 0.12
Cycle shift motif
Non cycle shift
Ultima
Illumina
Sequencing platform

FIG. 11B

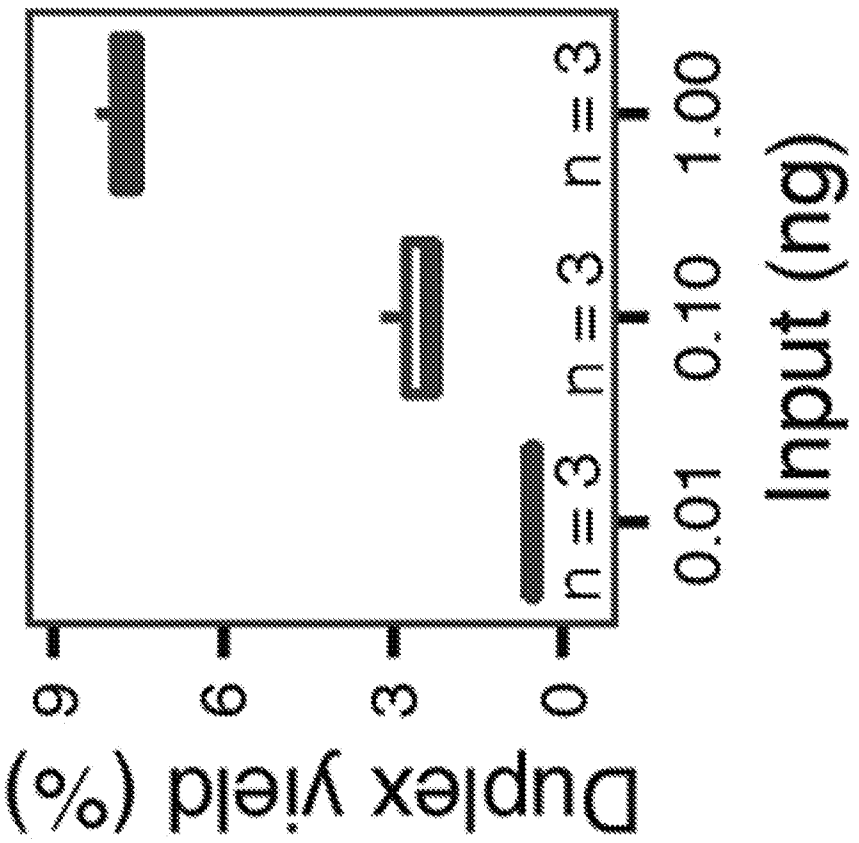


FIG. 12A

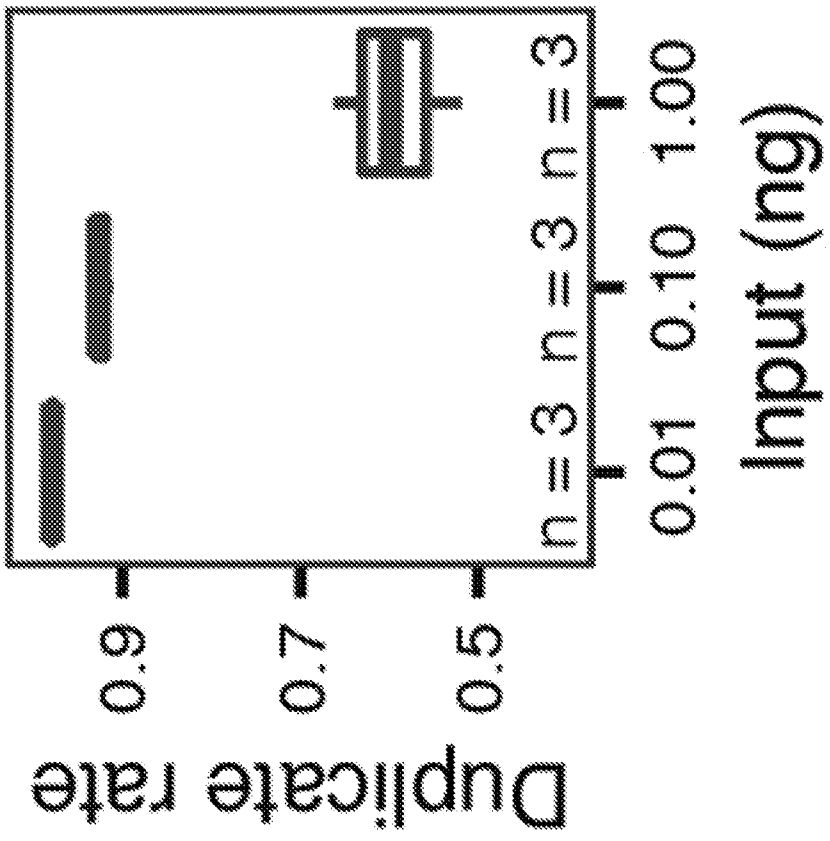


FIG. 12B

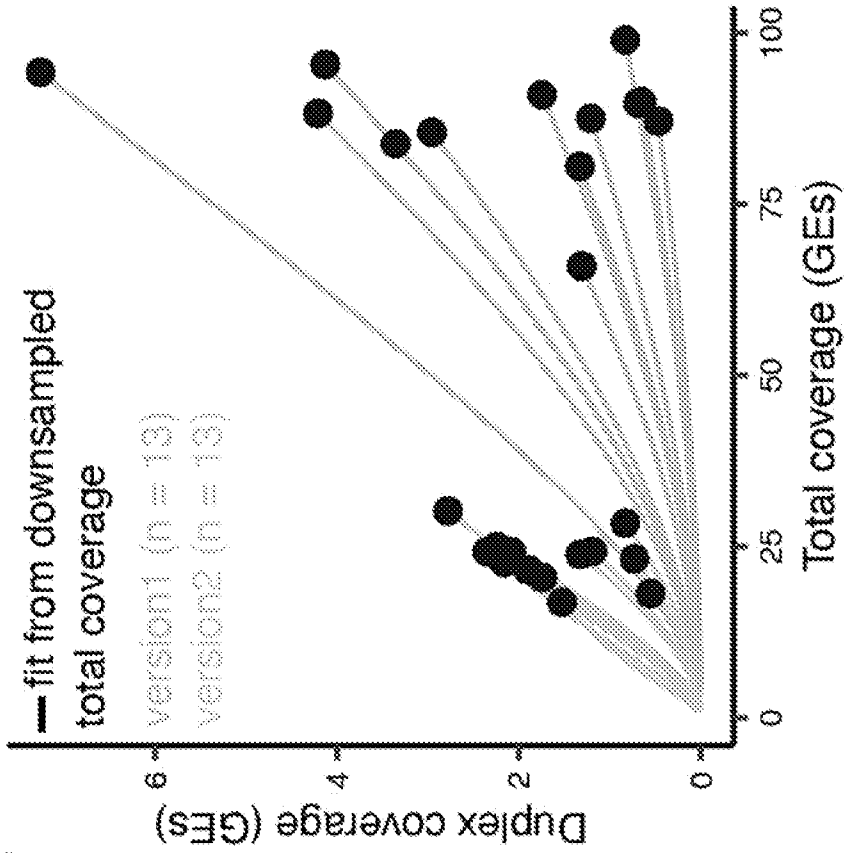

fit from downsampled total coverage
version1 (n = 13)
version2 (n = 13)
Duplex coverage (GEs)
0
2
4
6
0
25
50
75
100
Total coverage (GEs)

FIG. 12C

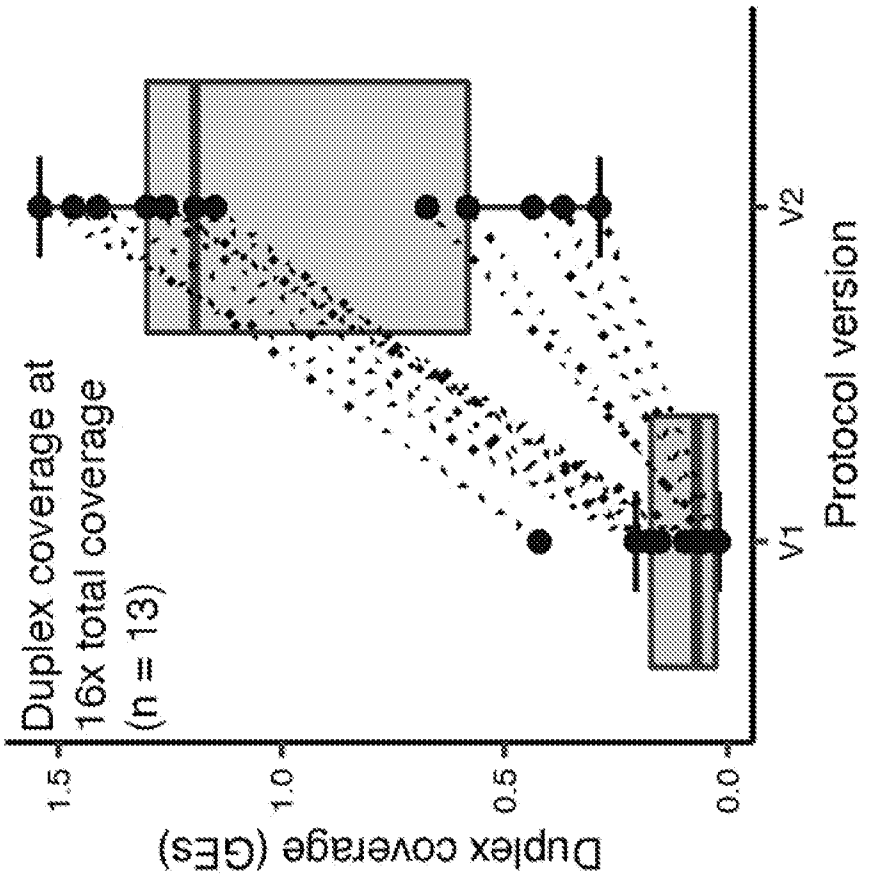

Duplex coverage at
16x total coverage
(n = 13)
Duplex coverage (GEs)
1.5
1.0
0.5
0.0
V1
V2
Protocol version

FIG. 12D

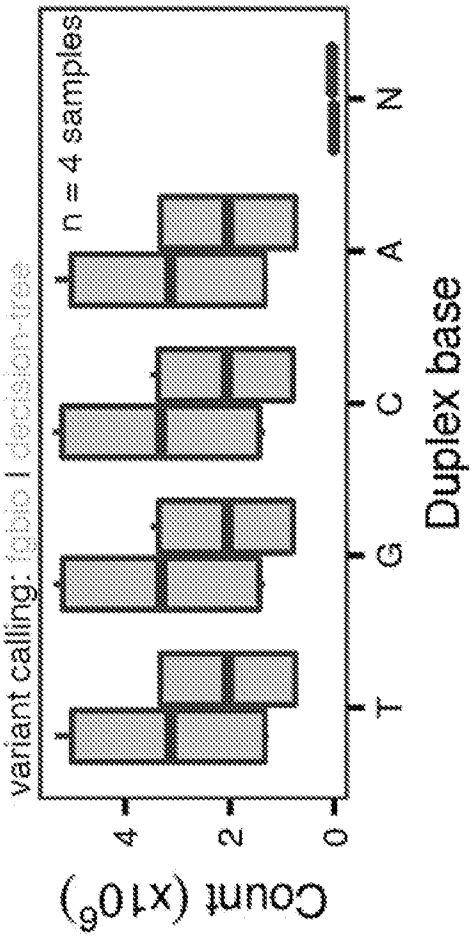

variant calling: fgbio | decision-tree
n = 4 samples
Count (x10^6)
0
2
4
T
G
C
A
N
Duplex base

FIG. 12E

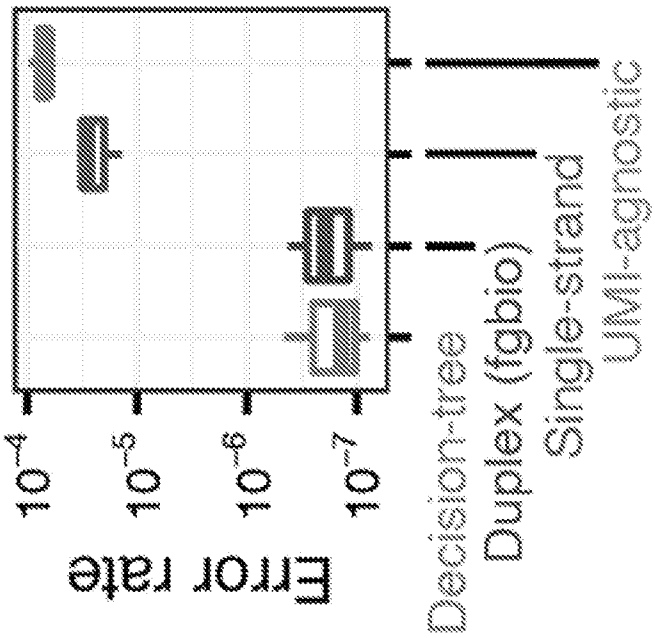

Error rate
10^{-4}
10^{-5}
10^{-6}
10^{-7}
Decision-tree
Duplex (fgbio)
Single-strand
UMI-agnostic

FIG. 13

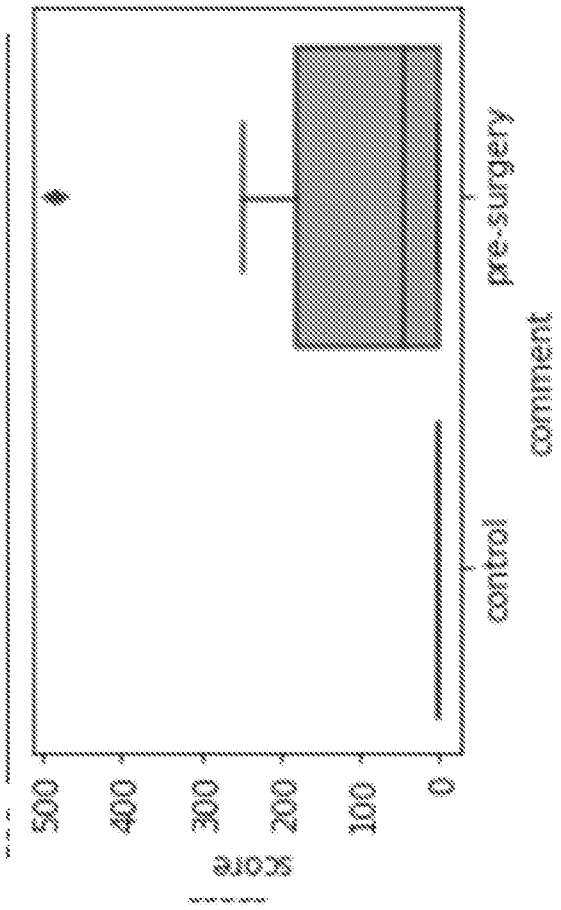


FIG. 14

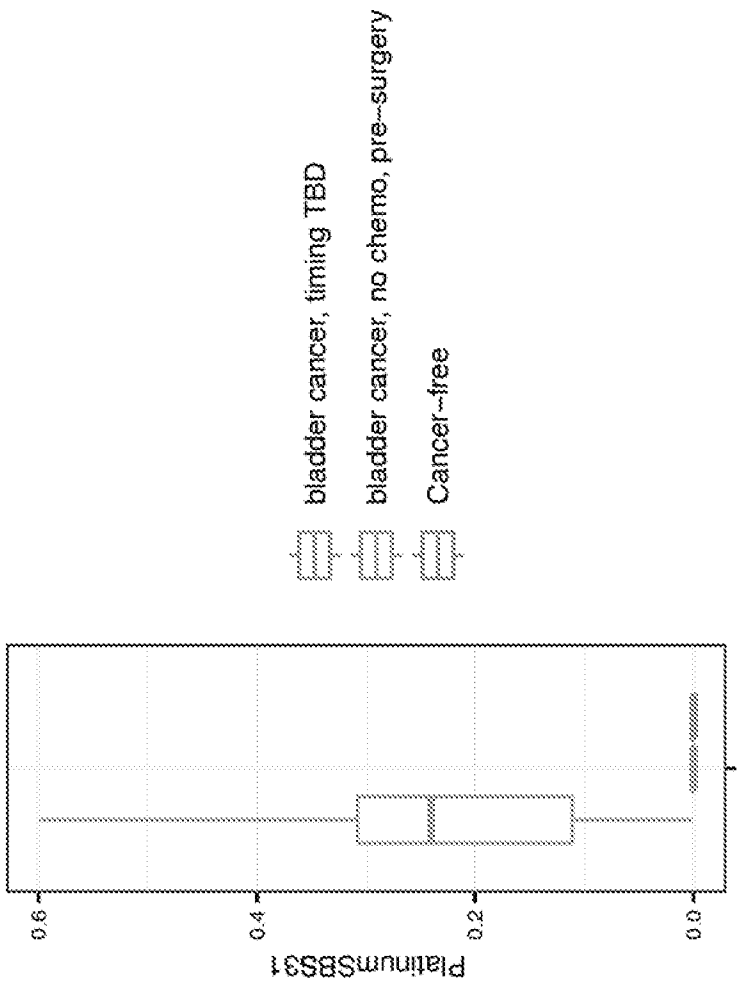

bladder cancer, timing TBD
bladder cancer, no chemo, pre–surgery
Cancer–free
PlatinumSBS31
0.6
0.4
0.2
0.0

FIG. 15

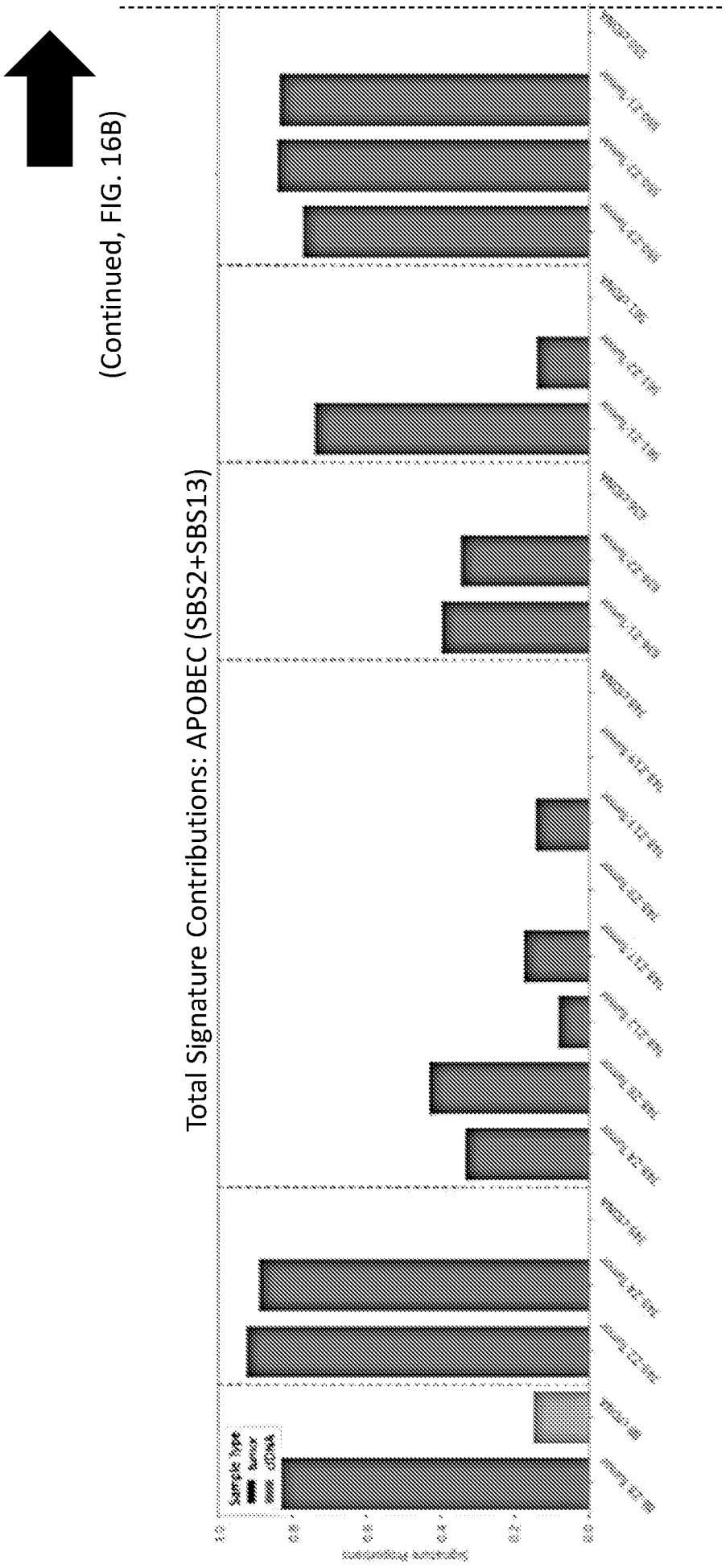
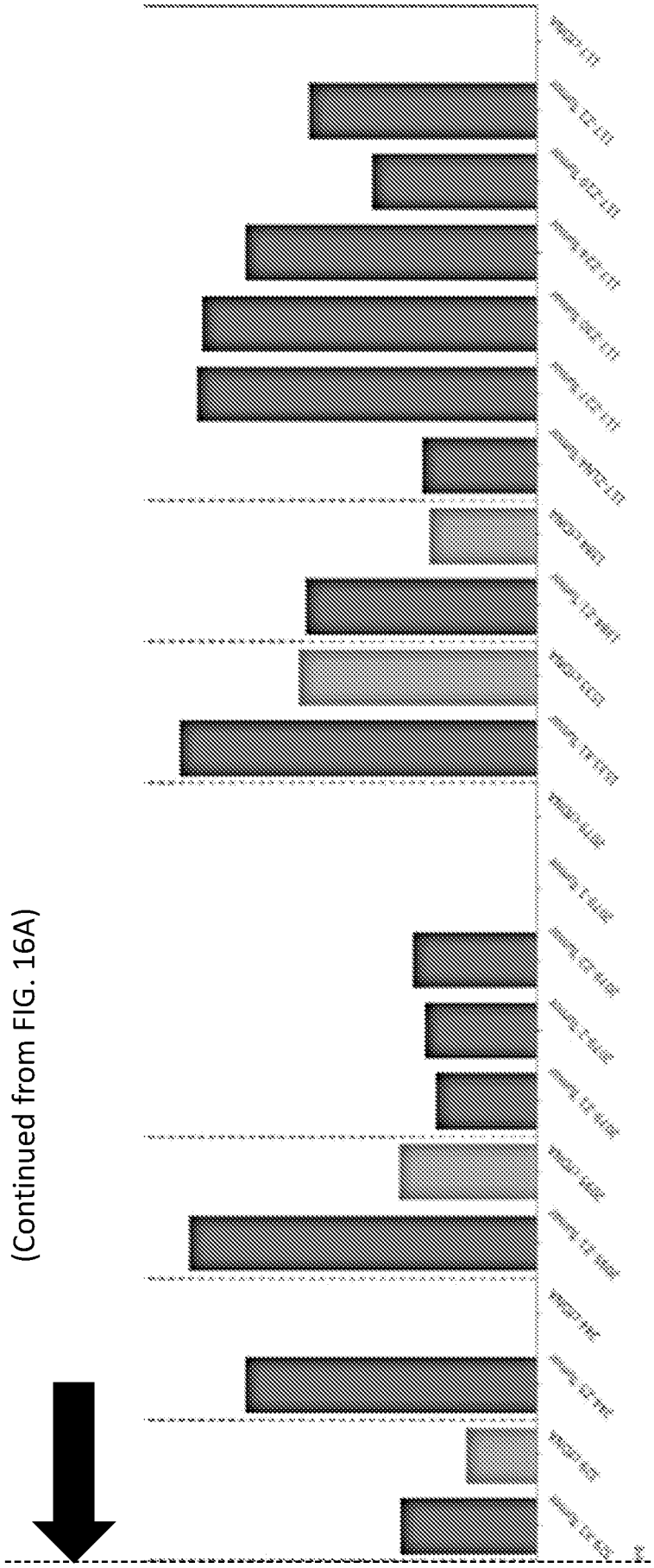

(Continued, FIG. 16B)
Total Signature Contributions: APOBEC (SBS2+SBS13)

FIG. 16A



(Continued from FIG. 16A)

FIG. 16B

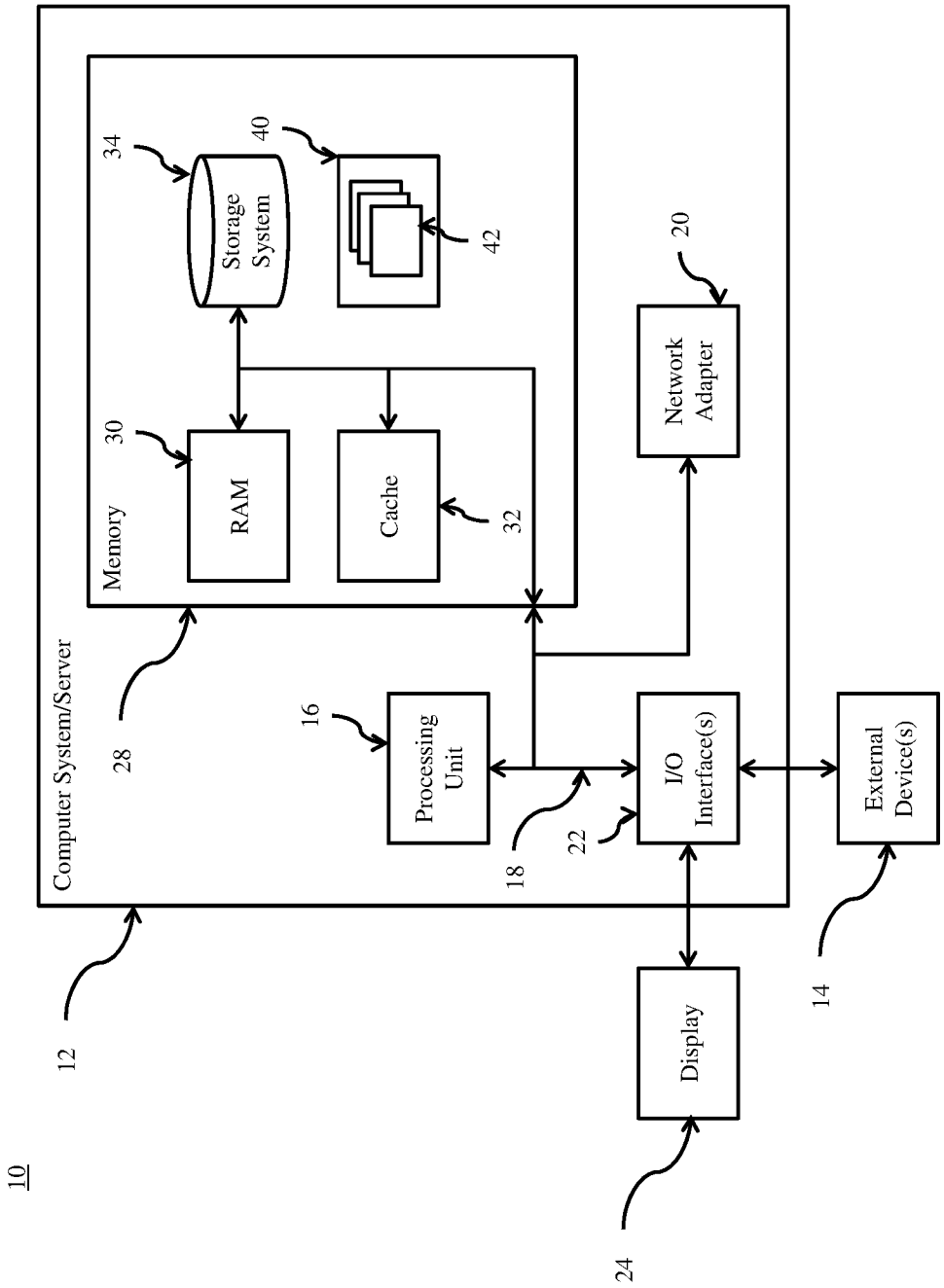

10
12
Computer System/Server
28
Memory
30
RAM
Cache
32
34
Storage System
40
42
16
Processing Unit
18
22
I/O Interface(s)
20
Network Adapter
External Device(s)
14
Display
24

FIG. 17