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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 October 2022 (13.10.2022)

WIPO | PCT

(10) International Publication Number
WO 2022/217096 A3

(51) International Patent Classification:
*G16B 30/20* (2019.01) *C12Q 1/6886* (2018.01)
*C12Q 1/6869* (2018.01)

(21) International Application Number:
PCT/US2022/024082

(22) International Filing Date:
08 April 2022 (08.04.2022)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/172,590 08 April 2021 (08.04.2021) US
63/276,378 05 November 2021 (05.11.2021) US

(72) Inventors; and
(71) Applicants: **HA, Gavin** [US/US]; c/o Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109 (US). **MACPHERSON, David** [US/US]; c/o Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109 (US). **NELSON, Peter S.** [US/US]; c/o Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109 (US). **DOEBLEY, Anna-Lisa** [US/US]; c/o Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109 (US). **HIATT, Joseph B.** [US/US]; c/o Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109 (US). **DE SARKAR, Navonil** [IN/US]; c/o Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109 (US). **PATTON, Robert** [US/US]; c/o Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109 (US).

(74) Agent: **SHELDON, David P.**; Christensen O'Connor Johnson Kindness PLLC, 1201 Third Avenue Suite 3600, Seattle, Washington 98101-3029 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, 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.

(54) **Title:** CELL-FREE DNA SEQUENCE DATA ANALYSIS METHOD TO EXAMINE NUCLEOSOME PROTECTION AND CHROMATIN ACCESSIBILITY

(57) **Abstract:** In one aspect, the disclosure provides a computer-implemented method of enhancing sequence read data from cell-free DNA samples for cell type prediction. The method comprises receiving sequence read data that includes a plurality of fragment reads, wherein each fragment read has a fragment length and a GC content indicating a percentage of bases in the fragment read that are G or C. GC bias values are determined a computing system for each fragment read based on the fragment length and the GC content of the fragment read. A genomic coverage distribution is generated that is adjusted for GC bias using the sequence read data and the GC bias values. Based on the genomic coverage distribution, the cell type is predicted. This method can be leveraged to assess cell subtypes and phenotypes based on cell free DNA present in biological samples for, e.g., cancer diagnosis, monitoring, and precision therapy.

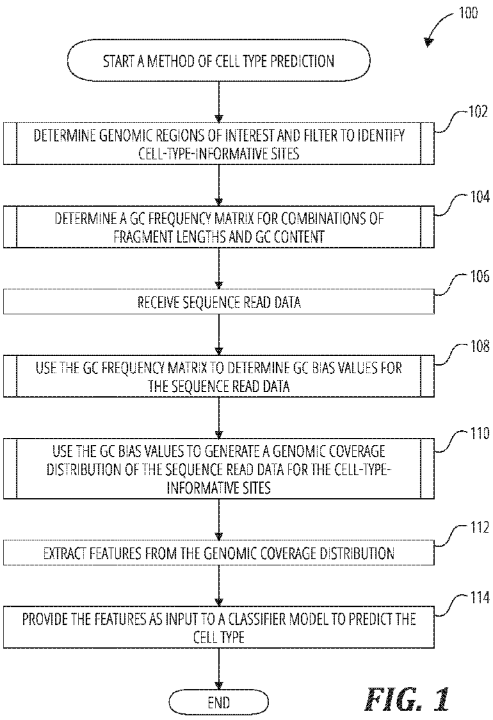

100
START A METHOD OF CELL TYPE PREDICTION
102
DETERMINE GENOMIC REGIONS OF INTEREST AND FILTER TO IDENTIFY CELL-TYPE-INFORMATIVE SITES
104
DETERMINE A GC FREQUENCY MATRIX FOR COMBINATIONS OF FRAGMENT LENGTHS AND GC CONTENT
106
RECEIVE SEQUENCE READ DATA
108
USE THE GC FREQUENCY MATRIX TO DETERMINE GC BIAS VALUES FOR THE SEQUENCE READ DATA
110
USE THE GC BIAS VALUES TO GENERATE A GENOMIC COVERAGE DISTRIBUTION OF THE SEQUENCE READ DATA FOR THE CELL-TYPE-INFORMATIVE SITES
112
EXTRACT FEATURES FROM THE GENOMIC COVERAGE DISTRIBUTION
114
PROVIDE THE FEATURES AS INPUT TO A CLASSIFIER MODEL TO PREDICT THE CELL TYPE
END

*FIG. 1*

WO 2022/217096 A3

[Continued on next page]

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

**Published:**

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

**(88) Date of publication of the international search report:**
29 December 2022 (29.12.2022)

# CELL-FREE DNA SEQUENCE DATA ANALYSIS METHOD TO EXAMINE NUCLEOSOME PROTECTION AND CHROMATIN ACCESSIBILITY

## CROSS-REFERENCES TO RELATED APPLICATIONS

This application claims the benefit of U.S. Patent Application No. 63/172,590, filed April 8, 2021, and U.S. Patent Application No. 63/276,378, filed November 5, 2021, the disclosures of which are incorporated herein by reference in their entireties.

## STATEMENT OF GOVERNMENT LICENSE RIGHTS

This invention was made with Government support under CA228944, CA264383, CA237746, CA097186, CA234715, CA076930, and HL007093 awarded by the National Institutes of Health and W81XWH-21-1-0513, W81XWH-18-1-0406, and W81XWH-17-1-0380 awarded by the United States Army Medical Research and Development Command. The Government has certain rights in the invention.

## BACKGROUND

Metastatic cancer is a late stage of cancer that often leads to cancer-related deaths. At the time of a *de novo* or recurrent metastatic cancer diagnosis, treatment options are often based on clinical diagnostics from the primary tumor. Additionally, molecular changes in the tumor, such as genetic alterations or phenotype changes, can emerge during metastatic progression or the development of treatment resistance. For instance, hormone receptor conversions in breast cancer, are frequent observed during the development of targeted treatment resistance. Therefore, it is important to classify tumor subtypes and identify patterns of transcriptional regulation that drive tumor phenotype changes during therapy. This type of work has critical implications for studying mechanisms of resistance to therapies and informing clinical treatment decisions in order to provide patients with life-prolonging treatment and care.

Current approaches for subtyping for solid tumors involve collecting a tissue biopsy and applying imaging techniques, such as immunohistochemistry, to assess the cellular phenotype. However, surgical biopsies for disease monitoring are often difficult to obtain from patients with metastatic cancer due to location and/or number of metastatic sites, especially in late-stage cancer. Moreover, surveillance of molecular changes in tumors is especially challenging because repeated biopsies are intractable and not considered standard-of-care. Thus, while accurate subtype determination is critical to address subtype switches during tumor recurrence and treatment resistance, tumor evolution and subtype plasticity during therapy is difficult to characterize, exemplifying a major limitation of current treatment strategies and precision medicine for patients with metastatic cancer.

To illustrate, breast cancer is among the most common causes of cancer, accounting for 23% of cancer diagnoses and 14% of cancer-related deaths among women worldwide. Targeted therapy is guided by tumor subtype, including the expression of three hormone receptors: ER, PR and HER2. In approximately 15% of cases, breast cancer tumors will undergo a switch in hormone subtype during tumor recurrence or as a mechanism of resistance to endocrine therapy. However, clinical determination of tumor subtype remains restricted to use of tissue biopsies, which are not routinely collected in late-stage cancers or repeatedly taken during the course of therapy.

Similarly, prostate cancer is the second most common cause of cancer mortality among men with an estimated 33,000 deaths in the United States in 2020. Castration-resistant prostate cancer (CRPC) describes the stage in which the disease has developed resistance to androgen deprivation therapy and progression to metastatic CRPC (mCRPC), which is an invariably lethal stage with no curative treatment. mCRPC is recognized to comprise multiple distinct subtype lineages and molecular subtypes, which are generally classified by specific genomic or epigenetic modifications. Prostate cancer can be categorized by phenotypic features that includes spectrum of trans-differentiated disease state including neuroendocrine (NE) carcinomas, low androgen regulated disease state (ARlowPC), double negative prostate cancer (DNPC, AR negative NE negative). These phenotypic subtypes are increasingly frequent as they emerge in the context of potent androgen ablation therapy resistance and are particularly important as they indicate distinctive prognosis of patients with advanced prostate cancer. In the new era of prostate cancer precision medicine several newer therapeutic options have emerged which are guided by a specific molecular alteration in the tumor genome (PARP inhibitors, immune checkpoint blockades are now guided by specific DNA repair aberrations). Accurate molecular classification defined by relevant molecular states will be critically important as it will create scopes of future "precision medicine" in prostate cancer.

Accordingly, there remains a need for facile and accurate methods to differentiate cell or tissue types, such as cancer phenotypes, that can be routinely performed in the clinic. The present disclosure addresses these and related needs.

## SUMMARY

This summary is provided to introduce a selection of concepts in a simplified form that are further described below in the Detailed Description. This summary is not intended to identify key features of the claimed subject matter, nor is it intended to be used as an aid in determining the scope of the claimed subject matter.

In one aspect, the disclosure provides a computer-implemented method of enhancing sequence read data from cell-free DNA samples for cell type prediction. The method comprises:

receiving, by a computing system, sequence read data, wherein the sequence read data includes a plurality of fragment reads, wherein each fragment read has a fragment length and a GC content indicating a percentage of bases in the fragment read that are G or C;

determining, by the computing system, GC bias values for each fragment read based on the fragment length and the GC content of the fragment read;

generating, by the computing system, a genomic coverage distribution that is adjusted for GC bias using the sequence read data and the GC bias values; and

predicting, by the computing system, the cell type based on the genomic coverage distribution.

In one embodiment, predicting the cell type based on the genomic coverage distribution includes predicting a cell phenotype. In one embodiment, predicting the cell phenotype includes predicting a tissue type, a cancer type, or a cancer subtype. In one embodiment, predicting the cell phenotype includes predicting expression of one or more genes of interest. In one embodiment, determining the GC bias value based on the fragment length and the GC content of the fragment read includes: counting a number of observed reads of each combination of fragment length and GC content to determine GC counts for the sequence read data; dividing the GC counts by corresponding GC frequencies in a GC frequency matrix to determine a GC bias for each fragment length; normalizing a mean GC bias for each fragment length to determine rough GC bias values; and smoothing the rough GC bias values to determine the GC bias values. In one embodiment, the GC frequency matrix stores a frequency for each GC content for each fragment length of a plurality of fragment lengths in mappable regions of a reference genome. In one embodiment, the plurality of fragment lengths includes each fragment length from a short length threshold to a long length threshold. In one embodiment, the short length threshold is in a range of 10-20 base pairs, and the long length threshold is in a range of 450-550 base pairs. In one embodiment, the short length threshold is 15 base pairs, and the long length threshold is 500 base pairs. In one embodiment, the method further comprises: determining genomic regions of interest for a cell type; and filtering the genomic regions of interest to identify cell-type-informative sites. In one embodiment, determining the genomic regions of interest includes: determining a mean mappability in a fixed size window around each genomic region of interest; and discarding genomic regions of interest having a mean mappability less than a predetermined threshold. In one embodiment, filtering the genomic regions of interest to identify cell-type-informative sites includes determining sites that have differential signals between a first cell type and a second cell type.

In one embodiment, generating the genomic coverage distribution includes: determining fragment midpoints in a window around each cell-type-informative site; assigning a weight for each fragment read based on an inverse of the GC bias value for each fragment read; using the weighted fragment reads to determine GC-corrected midpoint coverage profiles; excluding positions that overlap excluded regions; determining a mean profile based on determining an average of GC-corrected midpoint coverage profiles for all sites; smoothing the mean profile to generate a smoothed mean profile; and normalizing the smoothed mean profile by dividing by a mean of surrounding coverage to determine a normalized mean profile. In one embodiment, the excluded regions include one or more regions that are within an encode unified GRCh38 exclusion list, centromeres, gaps in human genome assembly, fix patches, alternative haplotypes, regions of zero mappability, or have coverage of at least 10 standard deviations above a mean. In one embodiment, predicting the cell type based on the genomic coverage distribution includes: generating one or more features based on the genomic coverage distribution; providing the one or more features as input to a classifier model; and determining the cell type based on an output of the classifier model. In one embodiment, the one or more features include a mean of coverage in a first predetermined window around each cell-type-informative site, a mean of coverage in a second predetermined window of a different size than the first predetermined window around each cell-type-informative site, and an amplitude of the genomic coverage distribution around each cell-type-informative site. In one embodiment, the first predetermined window is larger than the second predetermined window. In one embodiment, the first predetermined window has a width in a range of 1800-2200 base pairs, and the second predetermined window has a width in a range of 40-80 base pairs. In one embodiment, the first predetermined window has a width of 2000 base pairs, and the second predetermined window has a width of 60 base pairs. In one embodiment, the amplitude of the genomic coverage distribution around each cell-type-informative site is determined by: trimming the genomic coverage distribution to a window that contains 10 peaks; performing a fast Fourier transform on the window of the genomic coverage distribution; and determining a magnitude of the 10th frequency. In one embodiment, the classifier model includes a logistic regression model, an artificial neural network, a decision tree, a support vector machine, or a Bayesian network.

In another aspect, the disclosure provides a method of determining a chromatin accessibility profile for a cell of interest from a sample comprising cell-free DNA derived from the cell of interest. The method comprises:

obtaining sequence read data from the cell-free DNA;

receiving, by a computing system, sequence read data, wherein the sequence read data includes a plurality of fragment reads, wherein each fragment read has a fragment length and a GC content indicating a percentage of bases in the fragment read that are G or C;

determining, by the computing system, GC bias values for each fragment read based on the fragment length and the GC content of the fragment read;

generating, by the computing system, a genomic coverage distribution that is adjusted for GC bias using the sequence read data and the GC bias values; and

determining the chromatin accessibility profile from the genomic coverage distribution.

In one embodiment, the method further comprises determining a phenotype of the cell of interest based on the chromatin occupancy profile. In one embodiment, determining the cell phenotype comprises determining a tissue type, a cancer type, a cancer subtype, a malignancy aggressiveness phenotype, and/or a drug responsivity phenotype. In one embodiment, the method further comprises performing one or more steps of the computer implemented method described herein.

In another aspect, the disclosure provides a method for determining a cell type of a cell of interest from a sample comprising cell-free DNA derived from the cell of interest. The method comprises:

obtaining sequence read data generated from the sample comprising cell-free DNA;

performing the computer-implemented method described herein; and

determining the cell type of the cell of interest based on the prediction provided by the computing system.

In one embodiment, determining the cell type comprises determining a cell phenotype. In one embodiment, determining the cell phenotype comprises determining a tissue type, a cancer type, a cancer subtype, a malignancy aggressiveness phenotype, and/or a drug responsivity phenotype. In one embodiment, determining the cell phenotype includes determining expression of one or more genes of interest.

In another aspect, the disclosure provides a method of detecting the presence of a cancer cell in a subject, comprising:

obtaining sequence read data generated from the sample comprising cell-free DNA obtained from the subject;

performing the computer-implemented method described herein; and

determining the presence of a cancer cell in the subject based on the prediction provided by the computing system.

In one embodiment, the method is performed a plurality of times over time, wherein the detected cancer cell(s) in the subject at each performance of the method are further characterized to determine a cancer subtype or phenotype of the detected cancer cell(s) based on the prediction provided by the computing system. In one embodiment, the method is performed a plurality of times over time, and the method further comprises detecting a change in phenotype of the detected cancer cell(s) over time. In one embodiment, the subject receives a cancer therapy between performances of the method, and the method further comprises determining the responsivity of the cancer cell(s) to the treatment.

In another aspect, the disclosure provides a method of determining a cancer subtype of a target cancer cell from a sample comprising cell-free DNA derived from the target cancer cell. The method comprises:

obtaining sequence read data generated from the sample comprising cell-free DNA;

performing the computer-implemented method recited in any one of claims 5 to 21; and determining the cell type of the originating cell based on the predicted cancer subtype provided by the computing system.

In one embodiment, the sample is obtained from a subject with cancer. In one embodiment, the cancer is characterized as metastatic breast cancer. In one embodiment, determining the cancer subtype comprises determining whether the cancer is ER+ versus ER-. In one embodiment, determining the cancer subtype comprises determining whether the cancer is PR+ versus PR-. In one embodiment, determining the cancer subtype comprises determining whether the cancer is HER2+ versus HER2-. In one embodiment, determining the cancer subtype comprises determining two or all of:

whether the cancer is ER+ versus ER-,

whether the cancer is PR+ versus PR-, and

whether the cancer is HER2+ versus HER2-.

In one embodiment, cancer is characterized as metastatic prostate cancer. In one embodiment, determining the cancer subtype comprises determining whether the cancer is AR+ (ARPC) versus AR-. In one embodiment, determining the cancer subtype comprises determining whether the cancer is ARPC versus AR-low. In one embodiment, determining the cancer subtype comprises determining whether the cancer has a neuroendocrine prostate cancer (NEPC) phenotype signature or not. In one embodiment, determining the cancer subtype comprises determining whether the cancer is amphicrine. In one embodiment, determining the cancer subtype comprises determining two or all of:

whether the cancer is AR+ (ARPC) or AR-,

whether the cancer is AR-low or ARPC,

whether the cancer has a neuroendocrine prostate cancer (NEPC) phenotype signature or not,

whether the cancer is AR-low or NEPC,

whether the cancer is amphicrine or ARPC or NEPC.

In one embodiment, the cancer is characterized as lung cancer. In one embodiment, determining the cancer subtype comprises determining whether the cancer is small cell lung cancer (SCLC) or non-small cell lung cancer (NSCLC). In one embodiment, the method further comprises determining whether the NSCLC is adenocarcinoma or squamous cell carcinoma. In one embodiment, the sequence read data is generated from a panel of genomic targets. In one embodiment, the panel of genomic targets comprises transcription factor binding sites (TFBSs) of one or more transcription factors associated with SCLC. In one embodiment, the one or more transcription factors associated with SCLC comprise one or more of ASLC, NEUROD1, POU2F3, REST, and the like, and the method comprises determining the nucleosome occupancy of the TFBSs. In one embodiment, the TFBSs are identified by ChIP-seq data, or the like, and are retained in the panel if they are proximal to a transcription start site of a gene associated with lung cancer. In one embodiment, the panel of genomic targets comprise transcription start sites (TSSs) for one or more markers associated with lung cancer, wherein the method comprises determining the nucleosome occupancy of the TSSs.

In some embodiments of any method aspects described herein related to cancer detection or characterization, the sample is obtained from a subject. The method further can further comprise administering an effective treatment to the subject based on the determined cancer subtype. In one embodiment, the method further comprises performing the method on a plurality of samples obtained from the subject at a plurality of distinct time points after an initial diagnosis of cancer. In one embodiment, the sequence read data is generated by ultra-low pass whole genome sequencing. In one embodiment, sequence read data is generated by a chromatin accessibility assay. In one embodiment, sequence read data is generated in an ATAC-seq method. In one embodiment, sequence read data is generated in a ChIP-seq method. In one embodiment, sequence read data is generated in a DNAse sensitivity assay. In one embodiment, sequence read data is generated in a CUT&RUN assay. In one embodiment, CUT&RUN assay incorporates an affinity reagent that targets a post-translational modification to one or more of H3K27ac, H3K4me1 and H3K27ac.

In some embodiments of any method aspect described herein, the method can further comprises generating the sequence read data.

In some embodiments of any method aspect described herein, the sequence read data comprises sequence read data generated from a panel of genomic targets. In some embodiments of any method aspect described herein, the panel of genomic targets comprises transcription factor binding sites (TFBSs) of one or more transcription factors associated with a cancer type of interest. In some embodiments of any method aspect described herein, the method comprises determining the nucleosome occupancy of the TFBSs. In some embodiments of any method aspect described herein, the TFBSs are identified by ChIP-seq data, or the like, and are retained in the panel if they are proximal to a transcription start site of a gene associated with the cancer type of interest. In some embodiments of any method aspect described herein, the panel of genomic targets comprise transcription start sites (TSSs) for one or more markers associated with the cancer type of interest, wherein the method comprises determining the nucleosome occupancy of the TSSs.

In some embodiments of any method aspect described herein, the sample can be blood, plasma, or serum, and the like.

In another aspect, the disclosure provides a computer-implemented method of enhancing sequence read data from cell-free DNA samples for cell type prediction. The method comprises:

receiving, by a computing system, sequence read data, wherein the sequence read data includes a plurality of fragment reads, and wherein each fragment read has a fragment length;

determining, by the computing system, a fragment size variability for at least one gene associated with a cell type; and

predicting, by the computing system, the cell type based on the fragment size variability for the at least one gene.

In one embodiment, determining the fragment size variability includes determining a fragment size coefficient of variation. In one embodiment, predicting the cell type based on the genomic coverage distribution includes predicting a cell phenotype. In one embodiment, predicting the cell phenotype includes predicting a cancer subtype. In one embodiment, predicting the cell phenotype includes predicting a cancer subtype of prostate cancer. In one embodiment, predicting the cancer subtype includes distinguishing between ARPC and NEPC.

In one embodiment, predicting the cell type based on the fragment size variability includes:

generating one or more features based on the fragment size variability;

providing the one or more features as input to a classifier model; and

determining the cell type based on an output of the classifier model.

In one embodiment, generating the one or more features based on the fragment size variability includes generating a log2 fold change value of a fragment size coefficient of variation in a first cell type versus a second cell type. In one embodiment, the log2 fold change value predicts at least one of gene expression and gene transcriptional activity between the first cell type and the second cell type. In one embodiment, the first cell type is an ARPC cell and the second cell type is an NEPC cell. In one embodiment, the classifier model includes a logistic regression model, an artificial neural network, a decision tree, a support vector machine, or a Bayesian network.

In another aspect, the disclosure provides a method for determining a cell type of a cell of interest from a sample comprising cell-free DNA derived from the cell of interest, comprising:

obtaining sequence read data generated from the sample comprising cell-free DNA;

performing the computer-implemented method described herein (e.g., relating to predicting the cell type based on the fragment size variability); and

determining the cell type of the cell of interest based on the prediction provided by the computing system.

In one embodiment, determining the cell type comprises determining a cell phenotype. In one embodiment, determining the cell phenotype comprises determining a cancer subtype. In one embodiment, determining the cancer subtype includes distinguishing between ARPC and NEPC.

In another aspect, the disclosure provides a method of detecting the presence of a cancer cell in a subject, comprising:

obtaining sequence read data generated from a sample comprising cell-free DNA obtained from the subject;

performing the computer-implemented method described herein (e.g., relating to predicting the cell type based on the fragment size variability); and

determining the presence of a cancer cell in the subject based on the prediction provided by the computing system.

In one embodiment, the method is performed a plurality of times over time, wherein the detected cancer cell(s) in the subject at each performance of the method are further characterized to determine a cancer subtype or phenotype of the detected cancer cell(s) based on the prediction provided by the computing system. In one embodiment, the method is performed a plurality of times over time, and wherein the method further comprises detecting a change in phenotype of the detected cancer cell(s) over time. In one embodiment, the subject receives a cancer therapy between performances of the method, wherein the method further comprises determining the responsivity of the cancer cell(s) to the treatment.

In another aspect, the disclosure provides a method of determining a cancer subtype of a target cancer cell from a sample comprising cell-free DNA derived from the target cancer cell, the method comprising:

obtaining sequence read data generated from the sample comprising cell-free DNA;

performing the computer-implemented method described herein (e.g., relating to predicting the cell type based on the fragment size variability); and

determining the cell type of the originating cell based on the predicted cancer subtype provided by the computing system.

In one embodiment, the sample is obtained from a subject with cancer. In one embodiment, the cancer is characterized as metastatic prostate cancer. In one embodiment, determining the cancer subtype comprises determining whether the cancer is AR+ (ARPC) versus AR-. In one embodiment, determining the cancer subtype comprises determining whether the cancer is ARPC versus AR-low prostate cancer (ARLPC). In one embodiment, determining the cancer subtype comprises determining whether the cancer has a neuroendocrine prostate cancer (NEPC) phenotype signature or not. In one embodiment, the sample is obtained from a subject and the method further comprises administering an effective treatment to the subject based on the determined cancer subtype.

In one embodiment, the method further comprises performing the method on a plurality of samples obtained from the subject at a plurality of distinct time points after an initial diagnosis of cancer.

In one embodiment, the sequence read data is generated by ultra-low pass whole genome sequencing. In one embodiment, the sequence read data is generated by a chromatin accessibility assay. In one embodiment, the sequence read data is generated in an ATAC-seq method. In one embodiment, the sequence read data is generated in a ChIP-seq method. In one embodiment, the sequence read data is generated in a DNAse sensitivity assay. In one embodiment, the sequence read data is generated in a CUT&RUN assay. In one embodiment, the CUT&RUN assay incorporates an affinity reagent that targets a post-translational modification to one or more of H3K27ac, H3K4me1 and H3K27ac. In one embodiment, the method further comprises generating the sequence read data.

In one embodiment, the sequence read data is generated from a panel of genomic targets. In one embodiment, the panel of genomic targets comprises transcription factor binding sites (TFBSs) of one or more transcription factors associated with a cancer type of interest. In one embodiment, the method comprises determining the nucleosome occupancy of the TFBSs. In one embodiment, TFBSs are identified by ChIP-seq data, or the like, and are retained in the panel if they are proximal to a transcription start site of a gene associated with the cancer type of interest. In one embodiment, the panel of genomic targets comprise transcription start sites (TSSs) for one or more markers associated with the cancer type of interest, wherein the method comprises determining the nucleosome occupancy of the TSSs. In one embodiment, the sample is blood, plasma, or serum.

## DESCRIPTION OF THE DRAWINGS

The foregoing aspects and many of the attendant advantages of this invention will become more readily appreciated as the same become better understood by reference to the following detailed description, when taken in conjunction with the accompanying drawings, wherein:

FIGURE 1 is a flowchart that illustrates a non-limiting example embodiment of a method of cancer subtype prediction according to various aspects of the present disclosure.

FIGURE 2 is a flowchart that illustrates a non-limiting example embodiment of a procedure for determining informative sites for tissue, cell-type, cancer-type, or cancer-subtype of interest and filtering to identify cancer subtype-specific informative sites according to various aspects of the present disclosure.

FIGURE 3 is a flowchart that illustrates a non-limiting example embodiment of a procedure for determining a GC frequency matrix for a genome according to various aspects of the present disclosure.

FIGURE 4 is a flowchart that illustrates a non-limiting example embodiment of a procedure for using a GC frequency matrix to determine GC bias values for sequence read data according to various aspects of the present disclosure.

FIGURE 5 is a flowchart illustrating a non-limiting example embodiment of a procedure for using GC bias values to generate a nucleosome profile of sequence read data for subtype-specific informative sites according to various aspects of the present disclosure.

FIGURE 6 is a block diagram that illustrates aspects of an exemplary computing device appropriate for use as a computing device of the present disclosure.

FIGURES 7A and 7B illustrate the Griffin framework for cfDNA nucleosome profiling to predict cancer subtypes and tumor phenotype. FIGURE 7A is an illustration of a group of accessible sites (left panel) and inaccessible sites (right panel), such as a TFBS. The nucleosomes (in grey) are positioned in an organized manner around the accessible sites (box; left panel), but not around the inaccessible ones (right panel). These nucleosomes protect the DNA from degradation when it is released into peripheral blood. The protected fragments from the plasma are sequenced and aligned, leading to a coverage profile which reflects the nucleosome protection in the cells of origin. FIGURE 7B is a schematic showing the Griffin workflow for cfDNA nucleosome profiling analysis. cfDNA whole genome sequencing (WGS) data with ≥ 0.1x coverage is aligned to hg38 genome build. (1) For each sample, fragment-based GC bias is computed for each fragment size. (2) Sites of interest are selected from any assay. Paired-end reads aligned to each site are collected, fragment midpoint coverage is counted, and corrected for GC bias to produce a coverage profile. (3) Coverage profiles from all sites in a group (e.g., open chromatin for tumor subtype) are averaged to produce a composite coverage profile. Composite profiles are normalized using the surrounding region (-5 kb to +5 kb). (4) Three features are extracted from the composite coverage profile: central coverage (coverage from -30 bp to +30 bp from the site; 'a'), mean coverage (between -1 kb to +1 kb; 'b'), and amplitude calculated using a Fast-Fourier Transform (FFT) 'c').

FIGURES 8A to 8G illustrate that Griffin GC bias correction improves detection of tissue specific accessibility from cfDNA. FIGURE 8A graphically illustrates the aggregated GC content at 10,000 GRHL2 binding sites and its surrounding 2kb region. Mean GC content (line) and interquartile range (shading) are shown. FIGURE 8B graphically illustrates cfDNA GC bias is unique to each sample and each fragment length. GC bias computed for cfDNA from a healthy donor (HD_46; dashed shades) and a metastatic breast cancer (MBC_315; solid shades) sample are shown for various fragment sizes. FIGURE 8C graphically illustrates composite coverage profile of 10,000 GRHL2 binding sites before and after GC correction, shown for HD_46 (dashed) and MBC_315 (solid). Before GC correction, the 'central coverage' has a higher value due to effects of GC bias, which can obscure differential signals between samples. After GC correction, the central coverage of the MBC sample has lower value, which is consistent with increased GRHL2 activity in breast cancer but not immune cells making up the healthy donor sample. FIGURE 8D graphically illustrates composite coverage profiles of 10,000 LYL1 sites before and after GC correction, shown for two MBC samples with deep WGS (9-25x, orange), two healthy donors (17-20x, green), and 191 MBC samples with ULP-WGS (0.1-0.3x, blue). Median +/- IQR of 191 ULP-WGS samples is shown with blue shading. Lower 'central coverage' corresponding to greater site accessibility in the healthy donor samples is expected because LYL1 is a transcription factor associated with hematopoiesis. FIGURE 8E graphically illustrates cfDNA tumor fraction and central coverage correlation for LYL1, shown for ULP-WGS (0.1-0.3x, n=191) and WGS (9-25x, n=2) of MBC and healthy donors (17-20x, n=2) samples. cfDNA contains a mixture of tumor and blood cells; therefore, central coverage value is expected to be positively correlated with tumor fraction (lower represents increased accessibility). After GC correction, the correlation (for the MBC ULP-WGS samples) is much stronger based on Pearson's r correlation coefficient. Root mean squared error (RMSE) of the linear fit is shown. FIGURE 8F illustrates boxplots showing the distribution of the RMSE (linear fit between central coverage and tumor fraction in the MBC ULP-WGS dataset [0.1-0.3x, n=191]) across the 377 TFs, before and after GC correction. The boxed range represents the median ± IQR, whiskers represent the range of the non-outlier data (maximum extent is 1.5x the IQR). Outliers are plotted in grey. p-value was calculated using the Wilcoxon signed-rank test (two-sided). FIGURE 8G illustrates boxplots showing the distribution of the mean absolute deviation (of the central coverage across 215 healthy donors [1-2x WGS]) across the 377 TFs, before and after GC correction. Box elements are the same as (8F). p-value was calculated using the Wilcoxon signed-rank test (two-sided).

FIGURES 9A and 9B illustrate that Griffin enables accurate cancer detection and tissue-of-origin prediction. FIGURE 9A illustrates receiver operator characteristic (ROC) curve for logistic regression classification of cancer vs. healthy controls in three datasets, the DELFI dataset (Cristiano, S. et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature 570, 385–389 (2019)), LUCAS dataset, and LUCAS validation dataset (Mathios, D. et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. Nat Commun 12, 5060 (2021)). For each dataset, performance is shown for both the original low pass (1-2x) WGS and ultra-low pass (0.1x) WGS generated by in-silico downsampling. Logistic regression was performed on the top PCA components which explained 80% of the variance in the features (central coverage, mean coverage, and amplitude) extracted from nucleosome profiles around TFBSs. ROC for each stage of cancer vs. healthy are shown. FIGURE 9B illustrates boxplots of the AUC values for 1000 bootstrap iterations. The boxed range represents the median ± IQR, whiskers represent the range of the non-outlier data (maximum extent is 1.5x the IQR). Values below the boxplots show the median and 95% confidence interval.

FIGURES 10A to 10H illustrate that Griffin enables accurate prediction of breast cancer estrogen receptor subtypes from ultra-low pass WGS. FIGURE 10A: ER+ and ER- specific open chromatin sites were selected from assay for transposase-accessible chromatin using sequencing (ATAC-seq) data from ER+ (n=44) and ER- (n=15) breast tumors in The Cancer Genome Atlas (TCGA) (Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. Science 362, eaav1898 (2018)). Differential sites were identified using the DESeq2 software (Love, M.I., et al. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550 (2014)) to calculate the q-value and log2 fold change for each site. Sites with a q-value $<5*10^{-4}$ and a $\log_2$ fold change of >0.5 or <-0.5 were considered differential. FIGURE 10B illustrates composite coverage profiles (median ± IQR) for ER+ specific (n=18,240) and ER- specific (n=19,347) sites are shown for MBC patients (≥ 0.1 tumor fraction) separated by clinical ER status (ER+, n=50; ER-, n=51). Sites shared with hematopoietic cells were excluded from this figure (Satpathy, A. T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nature Biotechnology 37, 925–936 (2019)). FIGURE 10C illustrates a comut (Crowdis, J., He, M. X., Reardon, B. & Van Allen, E. M. CoMut: visualizing integrated molecular information with comutation plots. Bioinformatics 36, 4348–4349 (2020)) plot showing information about 101 MBC patients with >0.10 tumor fraction. Top row shows the ER status used for training and assessing the regression model. For most patients, this was the metastatic ER status obtained from IHC, if the metastatic ER status was not available, the primary ER status was used. ER low (1-10% ER staining) were considered ER positive. Second row, the upper left triangle contains the primary ER status and the lower right triangle contains the metastatic ER status. ER low biopsies are shown in light blue and unknown status is shown in white. Third row: tumor fraction, the fraction of cfDNA originating from the tumor, calculated using ichorCNA. Fourth row, median probability ER positive calculated by Griffin across 1000 bootstrap iterations. FIGURE 10D is a receiver operator characteristic (ROC) curve for a logistic regression model predicting ER+ and ER- subtype. ROC curve, accuracy and AUC are shown for all patients and for patients grouped by tumor fraction (TFx), 0.05-0.1 and ≥0.1. 95% CIs were obtained by bootstrapping. For patients with multiple samples, the first sample with tumor fraction >0.05 was used. FIGURE 10E graphically illustrates performance of the model on samples from three validation cohorts. For patients with multiple timepoints, the first sample was used. FIGURE 10F graphically illustrates subtype prediction in patients separated by clinical metastatic ER status and clinical primary tumor ER status. P-values were calculated using a Fisher's exact test (two-sided). FIGURE 10G illustrates ROC curve for predicting ER loss among patients with primary ER positive tumor. 95% CI was obtained by bootstrapping. FIGURE 10H illustrates the timeline for two patients (MBC 1413 and MBC 1099) with multiple biopsies of different subtypes and multiple cfDNA samples. ER+ prediction probability (thick grey line), and tumor fraction (thin grey line) is shown for all cfDNA samples that passed the >0.05 tumor fraction and 0.1x coverage thresholds. Decision boundary for ER+ (≥0.5) and ER- (<0.5) is indicated with dotted line. Timelines in months from metastatic diagnosis to death are shown for each patient. For patient MBC_1413, a metastatic biopsy (pleural fluid) was taken on the day of metastatic diagnosis and indicated ER- disease. However, approximately 7 months later, another metastatic biopsy (liver) showed weak ER+ staining (5%). A final biopsy (pleural fluid) taken at approximately 12 months and showed ER- staining once again. Plasma was drawn for cfDNA between the second and third metastatic biopsies with one final draw after then third biopsy. For patient MBC_1099 two ER- biopsies were taken at 0 months (bone) and 7 months (liver). cfDNA was drawn after this point, however between the two cfDNA draws, another biopsy (liver) indicated the presence of low level ER+ disease.

FIGURES 11A and 11B illustrate the workflow for characterizing advanced prostate cancer through matched tumor and liquid biopsies from PDX models. FIGURE 11A, top panel, illustrates that blood and tissue samples were taken from 26 patient-derived xenograft (PDX) mouse models with tumors originating from metastatic castration-resistant prostate cancer (mCRPC) with AR-positive adenocarcinoma (ARPC), neuroendocrine prostate carcinoma (NEPC) and AR-low non neuroendocrine prostate carcinoma (ARLPC) phenotypes. Cell-free DNA (cfDNA) was extracted from pooled plasma collected from 7-10 mice and whole genome sequencing (WGS) was performed. Following bioinformatic mouse read subtraction, pure human circulating tumor DNA (ctDNA) reads remained. From PDX tissue, ATAC-Seq and CUT&RUN (targeting H3K27ac, H3K4me1, and H3K27me3) data were generated. FIGURE 11A, middle panel, illustrates two distinct ctDNA features that were analyzed at transcription factor binding sites (TFBSs) and open chromatin sites throughout the genome using Griffin (see Example 1 and Doebley et al. (2021). Griffin: Framework for clinical cancer subtyping from nucleosome profiling of cell-free DNA. MedRxiv 2021.08.31.21262867 and Methods). FIGURE 11A, bottom right panel, shows phenotype classification using a probabilistic model that accounted for ctDNA tumor content and informed by PDX features was applied to 159 samples in three patient cohorts. FIGURE 11B illustrates PDX phenotypes and mouse plasma sequencing. Inclusion status based on final mean depth after mouse read subtraction (< 3x coverage were excluded unless AR coordinate amplification signal was reliably detected; lower dotted line). Phenotype status, including 6 NEPC, 18 ARPC (2 excluded), and 2 ARLPC. Average depth of coverage before and after mouse subtraction (mean coverage 20.5x; upper dotted line). Percentage of the cfDNA sample that contains human ctDNA after mouse read subtraction.

FIGURES 12A to 12G illustrate the analysis of tumor histone modifications and ctDNA reveals nucleosome patterns consistent with transcriptional regulation in CRPC phenotype-specific genes. FIGURE 12A illustrates H3K27ac peak signals between ARLPC, ARPC, and NEPC PDX tumor phenotypes at 10,000 AR binding sites (left) and at ASCL1 binding sites (right). Binding sites were selected from the GTRD (Yevshin et al. (2019). GTRD: a database on gene transcription regulation—2019 update. Nucleic Acids Res *47*, D100–D105) (Methods). FIGURES 12B and 12C graphically illustrate composite coverage profiles at 1000 AR (12B) and ASCL1 (12C) binding sites in ctDNA analyzed using Griffin. Coverage profile means (lines) and 95% confidence interval with 1000 bootstraps (shading) are shown. The region ±150 bp is indicated with vertical dotted line and yellow shading. FIGURE 12D is a heatmap of $\log_2$ fold change in key genes up and down regulated between ARPC and NEPC established through RNA-Seq (left) grouped by the type of histone modification which dictates translation levels: Group 1 shows genes where the predominate PTM mark is attributed to H3K27ac or H3K4me1 active marks in the gene promoters or putative distal enhancers, lacking H3K27me3 heterochromatic mark in the gene body; Group 2 features gene body spanning H3K27me3 repression marks. Central columns show differential peak intensity for each of the assayed histone modifications, separated by whether they appear upstream or in the promoter or the body of each gene. On the right the $\log_2$ fold change between ARPC and NEPC lines' fragment size coefficient of variation (CV) is shown for TSS+/- 1KB windows and respective gene bodies. FIGURE 12E graphically illustrates a comparison of the $\log_2$ fold change (ARPC vs. NEPC) of mean mRNA expression vs mean coefficient of variation (CV) in the 47 phenotypic lineage marker genes' promoter regions. FIGURE 12F (top) provides illustrations of expected ctDNA coverage profiles for Group 1 genes with and without H3K27ac or H3K4me1 modification leading to active and inactive transcription, respectively. FIGURE 12F (bottom) ±1000 bp surrounding the promoter region for AR and ASCL1 in ARPC and NEPC. Shown are coverage profile means (lines) and 95% confidence interval with 1000 bootstraps (shading). Decreased coverage is reflective of increased nucleosome accessibility and thus increased transcription. Dotted line and yellow shading highlight the focal window around transcription start site (TSS) (TSS -230 bp to +170 bp). FIGURE 12G is an illustration of expected ctDNA coverage profiles for Group 2 genes with repressed transcription caused by H3K27me3 modifications in the gene body. Neuronal gene UNC13A has increased nucleosome phasing in ctDNA of ARPC samples compared to NEPC.

FIGURE 13 illustrates hierarchical clustering of the normalized composite central mean coverage at TFBSs from the Griffin analysis of ctDNA for 107 TFs in LuCaP PDX lines of ARPC (n=16), NEPC (n=6), and ARLPC (n=2) phenotypes. This list of TFs was initially selected as having differential expression between ARPC and NEPC from LuCaP PDX RNA-Seq analysis. Heatmap colors indicate increased accessibility (low values; lighter) and decreased accessibility (higher values; darker) in ctDNA. TFs with increased accessibility in NEPC samples ($\log_2$-fold-change > 0.05, Mann-Whitney U test $p < 0.05$) are indicated with red text; increased accessibility in ARPC (log2-fold-change < -0.05, p < 0.05) are indicated with blue text.

FIGURES 14A to 14G illustrate comprehensive evaluation of ctDNA features throughout the genome for CRPC phenotype classification in PDX models. FIGURE 14A illustrates a volcano plot of $\log_2$-fold change of ATAC-Seq peak intensity between 5 ARPC and 5 NEPC lines; the dotted line demarcates sites by q-value < 0.05. FIGURES 14B and 14C graphically illustrate composite coverage profiles at open chromatin sites specific to ARPC (14B) and NEPC (14C) PDX tumors analyzed by Griffin. Sites from (14A) were filtered for overlap with known TFBSs in 338 factors from

GTRD (Yevshin et al. (2019). Nucleic Acids Res *47*, D100–D105). Coverage profile means (lines) and 95% confidence interval with 1000 bootstraps (shading) are shown. The region ±150 bp is indicated with vertical dotted line and yellow shading. FIGURE 14D illustrates CAs of ctDNA features demonstrates grouping between ARPC and NEPC phenotypes: (Left Panel) Composite central coverage of TFBSs significant for 74 TFs with differential accessibility out of 338 factors between ARPC and NEPC. (Right Panel) Fragment size variability (coefficient of variation) at H3K4me1 histone modification sites (n=9,750). FIGURE 14E graphically illustrates performance of classifying ARPC vs NEPC PDX from ctDNA using supervised machine learning (XGBoost) in various region types (all genes, TFBSs, and open regions, Methods). Area under the receiver operating characteristic curve (AUC) with 95% confidence interval (100 repeats of stratified cross validation) is shown for performance of all feature types. FIGURE 14F is an example composite coverage profiles at open chromatin sites specific to ARPC (left) and NEPC (right) identified in 14B-14C. Simulated admixtures generated using ARPC mixed with healthy donor (HD) (left) and NEPC mixed with HD (right) are shown for varying tumor fractions. FIGURE 14G graphically illustrates performance for classification on admixtures samples using the probabilistic mixture model. Five ctDNA admixtures were generated for each phenotype from PDX lines, each at various sequencing coverages and tumor fractions. In total, 125 admixtures were evaluated. The mean AUC across the 5 admixtures is shown for each configuration.

FIGURES 15A to 15C illustrate accurate classification of NEPC phenotypes from plasma in three patient cohorts using a probabilistic model informed by PDX ctDNA features. FIGURE 15A graphically illustrates receiver operating characteristic (ROC) curve for 101 mCRPC patients (DFCI cohort I) with ultra-low-pass WGS (ULP-WGS) data. The optimal performance of 90.4% sensitivity (for predicting NEPC) and 97.5% specificity (for predicting ARPC) corresponding to a prediction score cutoff of 0.3314 is indicated with horizontal and vertical dotted lines, respectively. FIGURE 15B illustrates prediction scores for 11 plasma samples from seven patients (DFCI cohort II) with both WGS and ULP-WGS data. The 0.3314 score cutoff threshold (dotted line) was used for classifying NEPC and ARPC. Tumor fractions were estimated by ichorCNA from WGS data. Patients were treated for adenocarcinoma (ARPC) or had high PSA values. FIGURE 15C illustrates prediction scores for 47 plasma samples with clinical phenotypes comprising 26 ARPC, 5 NEPC, and 16 mixed or ambiguous phenotypes (triangles), including double-negative prostate cancer (DNPC). Scores are shown for WGS and ULP-WGS (0.1X) for the same ctDNA sample. The cutoff threshold of 0.3314 (dotted line) was used for classifying NEPC and ARPC. Tumor fractions were estimated by ichorCNA on the WGS data.

FIGURE 16 is a schematic of an integrated, non-invasive targeted sequencing assay based on cfDNA for detection of genetic mutations and prediction of key tumor epigenetic features in SCLC.

FIGURES 17A and 17B illustrate the detection of transcription factor (TF) expression in SCLC models using targeted sequencing of cfDNA. FIGURE 17A is a schematic of experimental workflow for proof-of-concept negative control ("healthy donor") and positive control ("flank tumors" from SCLC cellular models) samples. FIGURE 17B graphically illustrates aggregated coverage across TFBSs in targeted sequencing data for healthy donors (top row) and flank tumors (bottom row). The TFBS is expected to be located at position 0 on the x axis. Data are color-coded by expected TF expression. Healthy donor-derived cfDNA is expected to reflect REST expression but not ASCL1, NEUROD1, or POU2F3. In SCLC models, systematic differences in coverage distribution as a function of TF expression are apparent.

FIGURES 18A to 18C illustrate transcription factor activity inference using TFBS coverage distributions from SCLC patient samples with available matched tumor gene expression data. FIGURE 18A graphically illustrates aggregated coverage across TFBSs in targeted sequencing data for healthy donors (top row) and patients with SCLC (bottom row) for whom matched tumor tissue with gene expression data was available. Samples are color-coded by expected TF expression. Systematic differences in coverage distribution as a function of expected TF expression are again apparent. FIGURE 18B illustrates gene expression of key genes in selected patient samples displayed as a heatmap. Cells are color coded by Z-score and the inset text is the log2(TPM+1). FIGURE 18C illustrates peak to trough amplitude calculated from coverage distributions at TFBS in each patient sample displayed as a heatmap. The amplitude is displayed by color and also as inset text. Trough depth magnitude corresponds to gene expression of the key TFs in these bona fide SCLC patient samples.

FIGURE 19 is a series of graphs illustrating quantification of transcription factor binding site peak to trough amplitude sample types. Distribution of TFBS peak to trough amplitude calculated from aggregated coverage distributions according to expected ground truth of TF expression. Pdx samples labeled "not SCLC" are NSCLC pdx models. Patient samples labeled "not SCLC" are either samples from patients with NSCLC (n=11) or without a diagnosis of malignancy (n=4). ASCL1 site peak to trough amplitude is associated with both SCLC status and ASCL1 positivity, while NEUROD1 and POU2F3 peak to trough amplitude is associated only with TF positivity.

FIGURES 20A and 20B graphically illustrate gene expression inference using TSS coverage distributions in flank tumor positive control samples. FIGURE 20A illustrates TSS coverage distribution from targeted sequencing of cfDNA, grouped by gene expression quintile in SCLC flank tumor models (quintiles 1-5) and blood ("B", dark blue). Shown are 1,912 TSS corresponding to 1,213 genes, which were selected based on low expression in whole blood and correlation between TSS coverage distribution and gene expression. TSS coverage distribution varies systematically according to expression of the corresponding gene. FIGURE 20B illustrates receiver operating characteristic curves for prediction of gene expression as above or below a threshold value (shown for thresholds of 0.1, 0.5, 1.0, and 2.0), as inferred from the coverage distribution of the corresponding TSS. An estimator of gene expression was calculated from the TSS coverage profile as the magnitude of the difference of the average coverage depth at positions +130 and +145 relative to the TSS minus the average depth at positions -45, -30, and -15 (shown as a dotted line in 20A). The AUC of the ROC curve is shown in parentheses for each gene expression cutoff. TSS coverage distributions can be used to predict whether a gene is expressed above or below a certain value with good test characteristics in this preliminary analysis that is restricted to especially variable, and therefore challenging, genes.

FIGURES 21A to 21C are a series of graphs illustrating use of aggregated coverage profiles across large rationally selected subsets of the TSS panel for prediction of SCLC vs NSCLC status in lung cancer Pdx models and Patient samples. The graphs provide examples of aggregated TSS coverage distributions across gene TSSs selected for upregulation in NSCLC (n=396) and SCLC (n=1045) for three different samples: one healthy donor (21A), one NSCLC Pdx model (21B), and one SCLC Pdx model (21C). As shown overlayed on the NSCLC PDX model, an amplitude feature was calculated from each coverage distribution curve as the difference between the coverage at the -45 position and the +120 position relative to the TSS, facilitating comparison within and between samples.

FIGURES 22A and 22B are a series of graphs illustrating use of aggregated coverage profiles across large rationally selected subsets of the TSS panel for prediction of SCLC vs NSCLC status in lung cancer Pdx models (22A) and Patient samples (22B. Aggregate coverage of SCLC-specific gene TSS (y axis, n=1045) vs NSCLC-specific gene TSS (x axis, n=396) in plasma samples from lung cancer PDX samples (non-cancer control patients also shown for reference as "benign") or from lung cancer patients. An SCLC PDX that transdifferentiated from an adenocarcinoma is identified with a thick red line.

FIGURE 23 is a flowchart that illustrates a non-limiting example embodiment of a method of cell (e.g., cancer, e.g., prostate cancer) subtype prediction according to an aspect of the present disclosure.

## DETAILED DESCRIPTION

The present disclosure is based on the inventors' development of a facile and sensitive approach to assess the chromatin architecture from cell-free DNA (cfDNA), and to provide accurate signal to detect and differentiate cell and/or tissue phenotypes based on the determined chromatin architecture.

Cell-free DNA (cfDNA) is released from dying cells, including tumor cells, and can be isolated from peripheral blood for studying aspects of biology. In the context of oncology, circulating tumor DNA (ctDNA) released from tumor cells into the blood as a subtype of cfDNA. The presence of ctDNA presents an opportunity for non-invasive "liquid biopsy" solution for addressing challenges in tissue accessibility, as described above. Current research and clinical efforts have focused on the detection of genetic mutations in select cancer genes from ctDNA and have demonstrated potential for clinical utility. Sequencing analysis of ctDNA to detect genomic alterations have also served to classify some subset of tumors based on genetic differences. However, studying the tumor phenotype from ctDNA remains challenging and is still a nascent area of research.

The inventors and others have noted that in the bloodstream, cfDNA is protected from degradation by nucleosomes and other DNA binding proteins, leading to a coverage pattern that reflects the genomic organization in the cells-of-origin. The genomic organization includes patterns of chromatin accessibility and transcriptional regulation, which, in turn, drive the differential phenotypes of the cells of origin. Thus, cfDNA can provide a non-invasive route to identify tumor subtypes through the analysis of tumor phenotypes beyond the traditional analysis of genotype, which involves DNA alterations.

While there is an intriguing possibility to perform "multi-omic" profiling of the genome, chromatin accessibility, transcriptome, and transcriptional regulation from analysis of ctDNA samples, there are heretofore a lack of robust tools to predict these multi-omic profiles from ctDNA, particularly for non-invasive applications to identify genomic and phenotypic signature changes during disease progression and treatment resistance in metastatic cancers. For example, prior efforts have been made to demonstrate that the nucleosome occupancy at transcription start site (TSS) of actively transcribed genes can be used to predict the presence of transcription for individual genes (see Ulz P, et al. Inferring expressed genes by whole-genome sequencing of plasma DNA. Nat Genet. 2016; 48(10):1273-1278, incorporated herein by reference in its entirety.) However, this approach required high (>75%) tumor fraction or regions of somatic copy number amplification for predictions with any reliable accuracy. In another study, the same idea was used to assess nucleosome occupancy at transcription factor binding sites (TFBSs) to predict transcription factor (TF) activity (Ulz P, et al. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nat Commun. 2019;10(1):4666, incorporated herein by reference in its entirety). This study demonstrated that TFs showed different signals between adenocarcinoma and NEPC ctDNA samples. However, the approach did not account for nor integrate features such as local sequence bias (e.g. GC content), somatic CNAs, ctDNA fraction, tumor fraction, and cfDNA fragment sizes, in which shorter fragments are enriched in cancer patients compared to healthy donors. These factors have been observed to greatly influence the results and, thus, obscure the actual signal from the data. Thus, it unlikely that such extant approaches can be sufficiently sensitive and robust to function for data from ultra-low-pass whole genome sequencing data (ULP-WGS) (0.1x), which is more cost-effective making it the accessible sequencing option for clinical settings.

The inventors have addressed the shortcomings of the art to produce a facile, robust, and sensitive approach to detecting and differentiating cell phenotypes. As described in more detail below, the approach is based in part on a core method, called "Griffin", to examine nucleosome protection and chromatin accessibility by quantifying cfDNA fragments around accessible sites. Unlike previous methods, Griffin implements critical approaches to consider fragment length-based GC correction to remove GC biases that obscure signals, which is especially prevalent in ULP-WGS applications (e.g., as low as 0.1x coverage of WGS). This novel fragment size-aware GC-bias correction approach as implemented in Griffin helps to maximize signal-to-noise and optimizes the analysis of sequence data, such as ULP-WGS of cfDNA. As an initial proof of concept described in Example 1, the inventors applied the Griffin approach to samples of cfDNA Griffin achieved excellent performance for detecting tumor cfDNA in early-stage cancer patients (AUC=0.96). Next, the approach was applied to samples obtained from subjects with metastatic breast cancer (MBC) of different phenotypes (i.e., ER+ and ER-) and demonstrated that the accurate determination of nucleosome occupancy allows differentiation of the two phenotypes. Specifically, 254 samples were analyzed from 139 patients and ER subtype was predicted with high performance (AUC=0.89), leading to valuable insights about tumor heterogeneity. This demonstration that the use of cfDNA can be used to predicting hormone subtypes will have immediate clinical diagnostic applications for identifying the subtype and potential switching in a minimally invasive and cost-effective manner, which can drive appropriate therapy. In this particular embodiment, the Griffin approach can be used for detection of breast cancer during early-stage disease, when treatments can have the greatest efficacy and impact.

The Griffin method provides additional advantages. Griffin is flexible to analyze any region throughout the genome that may be informative for differential chromatin accessibility between cell/tissue/cancer phenotype settings. For example, key transcriptional factors distinguishing between tumor subtypes can be predicted using Griffin via the analysis at binding sites of these transcription factors. Furthermore, Griffin can be applied to a variety of input data developed different assay approaches to study chromatin architecture and accessibility, including ATAC-seq, ChIP-seq, transcription factor profiling data, CUT & RUN, and the like. Moreover, in sharp contrast to existing technologies, Griffin can address countless hypotheses by enabling the analysis multiple 'omics', such as the following:

- Gene expression prediction (equivalent to transcriptomics)
- Transcriptional regulation, e.g. transcription factor activity (gene regulation; regulomics)
- Activity of transcription factors during targeted therapies (gene regulation; regulomics)
- Chromatin accessibility (epigenetics)
- Chromatin modifications, e.g. H3K27 acetylation (epigenetics)

- Hematopoietic and immune cell profiling (immunology)

The Griffin approach is adaptable to existing ctDNA sequencing techniques and, thus, permits scalability, adaptability, and accessibility, even from ULP-WGS data, which is highly susceptible to bias and signal obfuscation. Major applications of the approach include tumor (subtype) classification, identification of mixed histologies/phenotypes, detection of potential subtype switches (transdifferentiation) during therapy in "real time", and prediction of biomarkers (e.g., ARv7 splice variant) that can signal therapy resistance.

As described in more detail below (e.g., in Example 2), the inventors harnessed circulating tumor DNA (ctDNA) to study tumor phenotypes by ascertaining nucleosome positioning patterns associated with transcription regulation. Whole genomes of ctDNA in mouse plasma from 24 patient-derived xenograft models of androgen receptor active (ARPC) and neuroendocrine (NEPC) prostate cancers were sequenced. Nucleosome patterns associated with transcriptional activity were reflected in ctDNA at regions of genes, promoters, histone modifications, transcription factor binding, and accessible chromatin. The activity of key transcriptional regulators from ctDNA, including AR, ASCL1, HOXB13, HNF4G, and NR3C1, that were associated with prostate cancer phenotypes were identified. A prediction model was subsequently designed that distinguished NEPC from ARPC in 159 plasma samples across three clinical cohorts with 97-100% sensitivity and 85-100% specificity. These results highlight the utility of ctDNA in conjunction with the Griffin workflow for studying molecular phenotypes and advancing diagnostics in precision medicine.

In accordance with the foregoing, in one aspect, the disclosure provides a computer-implemented method of enhancing sequence read data from cell-free DNA samples for cell type prediction. In this context, the phrase "cell type prediction" is used in a general sense to refer to predicting the identity of, or a characteristic of, a cell of origin (i.e., a cell contributing DNA in the cfDNA sample). For example, the characteristic can be a distinguishable phenotype compared to cells with a same or similar developmental lineage, including developmental lineages with a transformation event (i.e., for cancer cells). Alternatively, the characteristic can be a distinguishable developmental lineage compared to a distinct developmental lineage. As described in more detail below, the method encompasses predicting or differentiating among different cell lineages, different tissue types, different tissue subtypes, different cancer types, difference cancer subtypes (i.e., subtypes of the same cancer type), and the like. The only requirement is that the cell type, as broadly defined, be distinguishable by a unique nucleosome occupancy and/or chromatin accessibility profile.

The method comprises:

receiving, by a computing system, sequence read data, wherein the sequence read data includes a plurality of fragment reads, wherein each fragment read has a fragment length and a GC content indicating a percentage of bases in the fragment read that are G or C;

determining, by the computing system, GC bias values for each fragment read based on the fragment length and the GC content of the fragment read;

generating, by the computing system, a genomic coverage distribution that is adjusted for GC bias using the sequence read data and the GC bias values; and

predicting, by the computing system, the cell type based on the genomic coverage distribution.

FIG. 1 is a flowchart that illustrates a non-limiting example embodiment of a method of cell type prediction according to various aspects of the present disclosure. The method 100 includes use of the GRIFFIN techniques described elsewhere herein to enable meaningful features to be extracted from short nucleic acid sequences of cancer DNA obtained from sequencing of cell-free DNA fragments in a sample. The method 100 may be used for various different types of cell type prediction, including but not limited to tissue type prediction, cell type prediction, cancer type prediction, and cancer subtype prediction.

From a start block, the method 100 proceeds to subroutine block 102, where genomic regions of interest are determined and filtered to identify cell-type-informative sites. Any suitable technique for determining and filtering cell-type-informative sites may be used, and different techniques will likely be used for different types of cancer, different molecular subtypes of a cancer type, different tissues, different cell types, and different types of assays. One non-limiting example embodiment of a suitable procedure for determining and filtering cell-type-informative sites is illustrated in FIG. 2 and described in further detail below.

At subroutine block 104, a GC frequency matrix is determined for combinations of fragment lengths and GC content. For certain sequencing technologies, fragments having certain amounts of G and C bases ("GC content") will be overrepresented in the sequence read data. This bias is not constant, as fragments of different sizes will have different GC biases. Because sequence read data from cell-free DNA fragments typically includes short fragments of many different lengths, establishing a GC frequency matrix that specifies expected proportions of GC content for various different fragment lengths allows sequence read data to be properly corrected for the GC bias, and for meaningful signals to be obtained from sequence read data that would otherwise be too noisy. One non-limiting example technique for determining a GC frequency matrix is illustrated in FIG. 3 and described in further detail below.

One will recognize that the actions described with respect to subroutine block 102 and subroutine block 104 may be performed on reference genome data before obtaining a sample or sequence data to be analyzed.

At block 106, sequence read data is received. In some embodiments, the sequence read data represents sequence reads generated for a sample obtained from a subject. In some embodiments, the sequence read data may be obtained from an archive or other previously obtained sample.

At subroutine block 108, the GC frequency matrix is used to determine GC bias values for the sequence read data. Any suitable technique may be used in subroutine block 108, including but not limited to the non-limiting example illustrated in FIG. 4 and described in further detail below.

At subroutine block 110, the GC bias values are used to generate a genomic coverage distribution of the sequence read data for the cell-type-informative sites. Again, any suitable technique may be used in subroutine block 110, including but not limited to the non-limiting example illustrated in FIG. 5 and described in further detail below.

At block 112, features are extracted from the genomic coverage distribution. Any features suitable for use with a classifier model may be extracted, and may depend on the type of classifier model used, the assay that generated the sequence reads, and/or the cell type (e.g., type of cancer, cancer subtypes, tissue, or cell type) to be detected. As one non-limiting example, for estrogen receptor (ER) subtyping in breast cancer, three features may be extracted: mean coverage, central coverage, and amplitude.

Mean coverage may be extracted by determining the mean coverage in a window around an informative site. The window around the informative site for determining mean coverage may be any suitable size, including but not limited to a range from 1800-2200 bp (from +/- 900 bp to +/- 1100 bp). One non-limiting example of a suitable size for the window for determining mean coverage is 2000 bp (+/- 1000 bp).

Central coverage may be extracted by determining the mean coverage in a smaller window around the informative site. The window around the informative site for determining central coverage may be any suitable size, including but not limited to a range from 40-80 bp (from +/- 20 bp to +/- 40 bp). One non-limiting example of a suitable size for the window for determining mean coverage is 60 bp (+/- 30 bp).

Amplitude may be extracted by trimming the genomic coverage distribution to an area that includes a given number of peaks (such as an area of +/- 960 bp that contains 10 peaks), performing a fast Fourier transform, and taking the magnitude of a frequency based on the given number of peaks (e.g., the 10th frequency for the area that contains 10 peaks).

At block 114, the features are provided as input to a classifier model to predict the cell subtype. Any suitable classifier model may be used. In one non-limiting example embodiment, the classifier model may be a logistic regression model.

Once the cancer subtype is predicted by the classifier model, the method 100 then proceeds to an end block and terminates. Naturally, in some embodiments, further action may be taken once the cancer subtype is determined, including but not limited to an appropriate cancer diagnosis, identifying cancer subtype change or switch, recommending a new course of treatment, altering an existing course of treatment, or any other appropriate action.

One consideration and challenge in analyzing plasma from patients is the presence of cfDNA released by hematopoietic cells, which leads to a lower ctDNA fraction (i.e., tumor fraction). Furthermore, the small patient cohorts with available tumor phenotype information make supervised machine learning approaches suboptimal. Therefore, an unsupervised probabilistic model was developed to estimate the proportion of cell types contributing to an individual plasma sample. One advantage of this model is the explicit modeling of the ctDNA tumor fraction in patients.

The input into this model includes signals generated from patient-derived xenografts (PDXs). PDXs provide a resource that is ideal for studying the properties of ctDNA, developing new analytical tools, and validating both genetic and phenotypic features by comparison to matching tumors. Using estimates of ctDNA fraction and these input PDX signals, the model applies a statistical mixture model approach to estimate the mixture weight parameter that represents the proportion of cell types. The mixture weight parameter may be used as a prediction score to classify cell types, such as ARPC and NEPC, as discussed below in Example 2 and illustrated in FIG. 14-15. Other cell types, such as phenotypes and subtypes, can also be modeled and predicted using this framework.

FIG. 2 is a flowchart that illustrates a non-limiting example embodiment of a procedure for determining genomic regions of interest and filtering to identify cell-type-informative sites according to various aspects of the present disclosure. In some embodiments, the cell types of interest for which the cell-type-informative sites are determined and filtered are different cancer types, different cancer subtypes, different tissue types, or different cell types.

From a start block, the procedure 200 advances to block 202, where a list of sites likely to be informative in the cell type of interest is selected. Sites may be selected using available data, including but not limited to public research databases and repositories, published scientific and sequencing data. These data may be derived from assays, including but not limited to sequencing techniques for Assay for Transposase-Accessible Chromatin (ATACs-eq), micrococcal nuclease (MNase-seq), DNAse hypersensitivity sites, chromatin immunoprecipitation (ChIP-seq), cleavage under targets & release using nuclease (CUT&RUN). Sites from these data that distinguish between cell types (e.g., tissue-types, cell-types, cancer-types, or cancer subtypes) are selected using any suitable comparison, including but not limited to statistical hypothesis testing using two-group Mann-Whitney U (also called Wilcoxon rank-sum) tests or Student-t's tests and multi-group Kruskal-Wallis test or analysis of variance (ANOVA). Additional filtering may be performed using fold change between groups.

At optional block 204, a mean mappability score (metric representing the uniqueness of the genomic sequence) is determined in a fixed size window around each site likely to be informative, and at optional block 206, sites having a mean mappability score less than a predetermined threshold are discarded. By keeping only sites that are mappable, the analysis is limited to sites that are likely to be accurately represented in the sequence read data. Mappability may be determined based on reference data, such as the mappability score track from the UCSC genome browser. In some embodiments, the actions of optional block 204 and optional block 206 may not be performed.

At block 208, the remaining sites that are informative for determining cell type are identified. Any suitable technique may be used. For example, for breast cancer ER subtyping, The Cancer Genome Atlas (TCGA) ATAC seq data may be used to identify sites that have differential ATAC signal between ER positive samples and ER negative TCGA samples. To identify these sites, any suitable technique may be used. In some embodiments, a Mann Whitney U test may be used at each site followed by false discovery rate (FDR) correction using the Benjamini-Hochberg procedure, retaining all sites with an adjusted p-value (i.e. q-value) below 0.05. In some embodiments, ATAC seq read counts around each site may be provided as input to DESeq2 software, which may then identify differential sites and produce an adjusted fold change and FDR corrected p-value for each site.

In some embodiments, the sites may be further refined by examining the fold change and retaining all sites with a log2 fold change greater than 0.5 in the subtype of interest relative to the other subtype. For breast cancer ER subtyping, ER positive and ER negative sites may be separated into those that are shared with hematopoietic cells and those which are not shared with hematopoietic cells using a separate dataset of hematopoietic ChIP seq peaks to generate a total of four subtype-specific informative site lists.

The procedure 200 then proceeds to an end block and terminates.

FIG. 3 is a flowchart that illustrates a non-limiting example embodiment of a procedure for determining a GC frequency matrix for a genome according to various aspects of the present disclosure. The technique described in FIG. 3 is different from previous techniques, such as the approach described in Benjamini & Speed, 2012 and implemented in DeepTools (Ramírez, Dündar, Diehl, Grüning, & Manke, 2014), at least because the previous techniques did not compensate for fragments of different lengths, and were never shown to work for cell-free DNA sequencing data. In the procedure 300, a separate GC bias curve is determined for each different fragment length.

First, we examine all mappable regions of the genome (block 302). Then, for each fragment length (for-loop defined between for-loop start block 304 and for-loop end block 310), we count a number of times each GC content is observed within fragments of the fragment length in the mappable regions to determine GC frequencies for the genome (block 306), and we store the GC frequencies in the GC frequency matrix for the fragment length (block 308).

After the for-loop, the procedure 300 advances to an end block and terminates.

In some embodiments, a range of fragment lengths between a short length threshold and a long length threshold are analyzed in the procedure 300. In some embodiments, the short length threshold may be in a range of 10-20 bp, and the long length threshold may be in a range of 450-550 bp. In one particular non-limiting example embodiments, the short length threshold may be 15 bp, and the long length threshold may be 500 bp. The for-loop may operate on each fragment length between the short length threshold and the long length threshold.

FIG. 4 is a flowchart that illustrates a non-limiting example embodiment of a procedure for using a GC frequency matrix to determine GC bias values for sequence read data according to various aspects of the present disclosure.

At block 402, the number of observed reads of each fragment length and GC content are counted to determine GC counts for the sequence read data.

At block 404, the GC counts are divided by the values in the GC frequency matrix to determine GC bias for each fragment length.

At block 406, a mean GC bias is normalized for each fragment length to determine rough GC bias values. In some embodiments, the mean GC bias may be normalized to 1. This results in a rough GC bias value for every possible combination of fragment size and GC content.

At block 408, the rough GC bias values are smoothed to determine the GC bias values. In some embodiments, for each fragment size, all GC bias values for similar sized fragments (as a non-limiting example, for 165 bp fragments, fragments of sizes from 155 bp to 175 bp may be considered) may be determined. The GC bias values for the similar sized fragments may be sorted by GC content, and kernel smoothing may be performed by taking the median of the nearest neighbors to determine the GC bias values.

The procedure 400 then advances to an end block and terminates.

FIG. 5 is a flowchart illustrating a non-limiting example embodiment of a procedure for using GC bias values to generate a genomic coverage distribution of sequence read data for cell-type-specific informative sites according to various aspects of the present disclosure.

From a start block, the procedure 500 advances to block 502, where fragment midpoints in a window around each cell-type-specific informative site are determined.

At block 504, a weight is assigned to each fragment based on the appropriate GC bias value for the fragment length and GC content (i.e., the GC bias value for the fragment length and GC content determined at subroutine block 108, e.g., by procedure 400). The weight is then based on that appropriate GC bias value. In some embodiments, the weight may be the inverse of the GC bias value (1 / GC bias value). For instance, if 165 bp fragments with 60% GC content have a GC bias of 2.5 in a given sample (overrepresented relative to 165 bp fragments with other GC contents), a weight of 1/2.5 = 0.4 would be assigned to these fragments.

At block 506, the weights are used to determine GC-corrected midpoint profiles.

At block 508, positions are excluded that overlap excluded regions. The excluded regions may be determined using any suitable technique. In some embodiments, the excluded regions may be obtained from one or more excluded region lists. Excluded region lists may include, but are not limited to, an encode unified GRCh38 exclusion list, centromeres, gaps in the human genome assembly, fix patches, alternative haplotypes, regions of zero mappability, and regions with unusually high coverage (e.g., 10 standard deviations above the mean).

At block 510, GC-corrected midpoint profiles for all sites are averaged to determine a mean profile.

At block 512, the mean profile is smoothed to generate a smoothed mean profile. Any suitable technique for smoothing may be used. For example, in some embodiments, the mean profile may be smoothed using a Savitzky-Golay filter with a window length of 165 bp and a 3rd order polynomial.

At block 514, the smoothed mean profile is normalized by dividing by the mean of the surrounding coverage. In some embodiments, surrounding coverage in a range of 9,000-11,000 bp (+/- 4,500 bp to +/- 5,500 bp), such as 10,000 bp (+/- 5,000 bp) is considered for normalization. This allows samples with different depths of sequencing coverage to be compared.

The normalized mean profile may be used as the resulting genomic coverage distribution.

The procedure 500 then advances to and end block and terminates.

FIG. 6 is a block diagram that illustrates aspects of an exemplary computing device appropriate for use as a computing device of the present disclosure. The techniques described above, including but not limited to the techniques described in method 100, may be implemented in full or in part on one or more computing systems that include one or more computing devices such as computing device 600 that are communicatively coupled to each other.

The exemplary computing device 600 describes various elements that are common to many different types of computing devices, including but not limited to desktop computing devices, laptop computing devices, server computing devices, mobile computing devices, and computing devices that are part of a cloud computing system. While FIG. 6 is described with reference to a computing device that is implemented as a device on a network, the description below is applicable to servers, personal computers, mobile phones, smart phones, tablet computers, embedded computing devices, and other devices that may be used to implement portions of embodiments of the present disclosure. Some embodiments of a computing device may be implemented in or may include an application-specific integrated circuit (ASIC), a field-programmable gate array (FPGA), or other customized device. Moreover, those of ordinary skill in the art and others will recognize that the computing device 600 may be any one of any number of currently available or yet to be developed devices.

In its most basic configuration, the computing device 600 includes at least one processor 602 and a system memory 610 connected by a communication bus 608. Depending on the exact configuration and type of device, the system memory 610 may be volatile or nonvolatile memory, such as read only memory ("ROM"), random access memory ("RAM"), EEPROM, flash memory, or similar memory technology. Those of ordinary skill in the art and others will recognize that system memory 610 typically stores data and/or program modules that are immediately accessible to and/or currently being operated on by the processor 602. In this regard, the processor 602 may serve as a computational center of the computing device 600 by supporting the execution of instructions.

As further illustrated in FIG. 6, the computing device 600 may include a network interface 606 comprising one or more components for communicating with other devices over a network. Embodiments of the present disclosure may access basic services that utilize the network interface 606 to perform communications using common network protocols. The network interface 606 may also include a wireless network interface configured to communicate via one or more wireless communication protocols, such as Wi-Fi, 2G, 3G, LTE, WiMAX, Bluetooth, Bluetooth low energy, and/or the like. As will be appreciated by one of ordinary skill in the art, the network interface 606 illustrated in

FIG. 6 may represent one or more wireless interfaces or physical communication interfaces described and illustrated above with respect to particular components of the computing device 600.

In the exemplary embodiment depicted in FIG. 6, the computing device 600 also includes a storage medium 604. However, services may be accessed using a computing device that does not include means for persisting data to a local storage medium. Therefore, the storage medium 604 depicted in FIG. 6 is represented with a dashed line to indicate that the storage medium 604 is optional. In any event, the storage medium 604 may be volatile or nonvolatile, removable or nonremovable, implemented using any technology capable of storing information such as, but not limited to, a hard drive, solid state drive, CD ROM, DVD, or other disk storage, magnetic cassettes, magnetic tape, magnetic disk storage, and/or the like.

Suitable implementations of computing devices that include a processor 602, system memory 610, communication bus 608, storage medium 604, and network interface 606 are known and commercially available. For ease of illustration and because it is not important for an understanding of the claimed subject matter, FIG. 6 does not show some of the typical components of many computing devices. In this regard, the computing device 600 may include input devices, such as a keyboard, keypad, mouse, microphone, touch input device, touch screen, tablet, and/or the like. Such input devices may be coupled to the computing device 600 by wired or wireless connections including RF, infrared, serial, parallel, Bluetooth, Bluetooth low energy, USB, or other suitable connections protocols using wireless or physical connections. Similarly, the computing device 600 may also include output devices such as a display, speakers, printer, etc. Since these devices are well known in the art, they are not illustrated or described further herein.

As indicated, the computer-implemented method implementing the Griffin workflow is highly adaptable to different types of input data reflective of the chromatin architecture (e.g., nucleosome occupancy and chromatin accessibility). The method can be applied to various contexts of analyses depending on the source and character of the originating cells or tissues being analyzed. Accordingly, in another aspect, the disclosure provides a method of determining a chromatin accessibility profile for a cell of interest from a sample comprising cell-free DNA derived from the cell of interest. This method applies the Griffin data optimization workflow, described in more detail above, to determine a chromatin accessibility profile for a cell of interest. The method is flexible and permits input data obtained from a variety of sequencing and capture protocols. The method comprises:

obtaining sequence read data from the cell-free DNA;

receiving, by a computing system, sequence read data, wherein the sequence read data includes a plurality of fragment reads, wherein each fragment read has a fragment length and a GC content indicating a percentage of bases in the fragment read that are G or C;

determining, by the computing system, GC bias values for each fragment read based on the fragment length and the GC content of the fragment read;

generating, by the computing system, a genomic coverage distribution that is adjusted for GC bias using the sequence read data and the GC bias values; and

determining the chromatin accessibility profile from the genomic coverage distribution.

The method can further comprise determining a phenotype of the cell of interest based on the chromatin occupancy profile. For example, determinations of cell phenotype can include determining the tissue type of origin of the cell, determining if the cell is transformed (e.g., is cancerous or malignant), determining the cancer type or cancer subtype, determining a malignancy aggressiveness phenotype, and/ or determining a drug responsivity phenotype. The term malignancy aggressiveness phenotype refers to the relative aggressiveness of a transformed (e.g., cancer) cell in terms of rate of reproduction, migration, drug responsivity, and the like. The phenotype can be qualitative or can be assessed by various metrics to allow for quantitative comparison. The term "drug responsivity phenotype" refers to the relative responsivity (i.e., susceptibility or resistance) of a cancer cell to a cancer therapy. The metric can be quantitative or qualitative. These determinations can be made using various classifiers, described in more detail above, based on sequence data optimized by the Griffin workflow. Elements of the Griffin workflow and computer implemented method are described in more detail above and incorporated into the present aspect without limitation. Exemplary, nonlimiting implementations of the Griffin workflow and associated classifiers to subtype cancer cells with distinct phenotypes are provided in the Examples.

As indicated herein, the Griffin workflow enhances data from a variety of sequencing and capture platforms to provide profiles of nucleosome accessibility, and these profiles can provide highly accurate insight as to the nature of cells that contribute to the ctDNA present in biological samples. These insights enable detecting and characterizing cells that contribute to the ctDNA, including enabling the ability to detect cells of a certain type and/or differentiate cells between various subtypes. Thus, in a particular aspect, the disclosure also provides a method for determining or identifying a cell type of a cell of interest from a sample comprising cell-free DNA derived from the cell of interest. The method of this aspect comprises:

obtaining sequence read data generated from the sample comprising cell-free DNA;

performing the computer-implemented method described in more detail above (and which is incorporated into this aspect in all of its embodiments); and

determining or identifying the cell type of the cell of interest based on the prediction provided by the computing system. The determining step can be performed by any of a number of appropriate classifiers based on the data enhanced by the Griffin workflow. As above, the determining step can comprise determining a cell phenotype, such as determining tissue type, a cancer type, a cancer subtype, a malignancy aggressiveness phenotype, a drug responsivity phenotype, or expression (or expression level) of a gene of interest.

In yet another aspect, the disclosure provides a method for detecting the presence of a cancer cell in a subject. The method comprises:

obtaining sequence read data generated from the sample comprising cell-free DNA obtained from the subject;

performing the computer-implemented method described in more detail above (and which is incorporated into this aspect in all of its embodiments); and

determining the presence of a cancer cell in the subject based on the prediction provided by the computing system.

In some embodiments, the method is performed a plurality of times. Accordingly, the method can be a method of monitoring for the presence and/or identity of cancer in the subject. The cancer cell(s) detected in the subject at each performance of the method can be further characterized. For example, the cell(s) can be monitored over time using this method to determine a cancer subtype or phenotype of the detected cancer cell(s) based on the prediction provided by the computing system. In some embodiments, the method further comprises detecting a change in phenotype of the detected cancer cell(s) over time. For example, as described in more detail below certain cancer types can progress from one subtype to another during the course of disease. Cancer cells can evolve and essentially switch between characterized subtypes. These changes can be associate with changes in malignancy and/or responsivity to various treatments, all of which can be detected given the demonstrated sensitivity of the Griffin workflow. Monitoring and documenting such changes over time can inform a requirement for modification of therapy to optimize the outcome. As a non-limiting example, non-small cell lung cancer (NSCLC) can be monitored for transdifferentiation to small cell lung cancer (SCLC). Alternatively, SCLC subtypes can be monitored for transdifferentiation to distinct subtypes. In some embodiments, the method can be performed starting before or during the course of treatment for cancer. Accordingly, the cancer can be monitored for responsivity to the treatment, or for changes in phenotype during the course of treatment. These characteristics can inform any appropriate adjustments to the treatment regimen. In some embodiments, the method comprises implementing a treatment or treatment change based on the monitored status of the cancer cells as determined by the method.

In another aspect, the disclosure provides a method of determining a cancer subtype of a target cancer cell from a sample comprising cell-free DNA derived from the target cancer cell. The method comprises:

obtaining sequence read data generated from the sample comprising cell-free DNA;

performing the computer-implemented method described in more detail above (and which is incorporated into this aspect in all of its embodiments); and

determining the cell type of the target cancer cell based on the predicted cancer subtype provided by the computing system.

The sample can be a biological sample from the subject, e.g., a subject with cancer or suspected to have cancer. Exemplary biological samples are described in more detail below. In some embodiments, the method comprises obtaining the biological sample from the subject and/or generating the sequence read data from the sample, according to standard techniques appropriate for the desired sequencing platform and/or targeted capture technology.

As described in more detail below, the Griffin platform has been employed to successfully distinguish between important subtypes of cancers for various different, unrelated cancers, indicating the broad applicability to cancer types in general. Thus, in some embodiments, the cancer is characterized as metastatic breast cancer. In some further embodiments, the determining step comprises determining the status of the breast cancer as ER+ versus ER-, which refers to the expression of estrogen receptor (ER) and whether the cancer cells respond to exposure of the estrogen hormone. This status can be a critical to inform the appropriate course of therapy because ER+ breast cancers can be addressed by administration of endocrine therapies. In other embodiments, the determining step comprises determining the status of the breast cancer as PR + versus PR -, which refers to the expression of progesterone receptor (PR) and whether the cancer cells respond to exposure of the progesterone hormone. Similarly, this status can be a critical to inform the appropriate course of therapy because PR+ breast cancers can also be addressed by administration of appropriate hormonal therapies, such as tamoxifen and aromatase inhibitors. In yet another embodiment, the determining step comprises determining the status of the breast cancer as HER2+ versus HER2-, which refers to the expression of human epidermal growth factor receptor 2 (HER2). HER2+ breast cancer cells tend to result in poorer prognosis as they grow faster and have a higher likelihood of spreading, e.g., to the lymph nodes. This status can be a critical to inform the appropriate course of therapy because PR+ breast cancers can also be addressed by administration of appropriate Her2-targeted therapy, such as trastuzumab or pertuzumab. Of course, it will be appreciated that the disclosure also encompasses embodiments of distinguishing determining the expression status of multiple informative markers. For example the method can comprise determining: whether the cancer is ER+ versus ER-; whether the cancer is PR+ versus PR-; and/or whether the cancer is HER2+ versus HER2-, in any combination. For example, in one embodiments, the method comprises determining whether the cancer is ER+ versus ER-, whether the cancer is PR+ versus PR-, and whether the cancer is HER2+ versus HER2-. Patients with triple-negative breast cancer (i.e. ER-, PR-, HER-) may receive neoadjuvant chemotherapy, such as carboplatin and paclitaxel, and in combination with immunotherapy, such as Pembrolizumab and atezolizumab.

In another embodiment, the cancer is characterized as metastatic prostate cancer. In a further embodiment, determining the subtype of the prostate cancer addresses determining whether the cancer expresses various markers characteristic of distinguishable subtypes. For example, in one embodiment, the step of the cancer subtype comprises determining whether the prostate cancer is AR+ (ARPC) versus AR-, which refers to the status for expression of androgen receptors. Alternatively, the step of the cancer subtype comprises determining whether the prostate cancer is AR+ (ARPC) versus AR (low). Prostate cancers that are AR+ are often treated with androgen receptor signaling inhibitors (ARSI) that repress the androgen receptor activity in the cells. In another embodiment, the step of the cancer subtype comprises determining whether the prostate cancer has a neuroendocrine prostate cancer (NEPC) phenotype signature or not. NEPC cells lack AR activity and possess distinct transcriptional programming regulation profiles from CRPC cells, including different epigenetic modifications, that result in a distinct phenotype that requires alternative therapeutic intervention. In another embodiment, the step of the cancer subtype comprises determining whether the prostate cancer is amphicrine, which refers to possessing both exocrine and neuroendocrine characteristics in the same cell. As is demonstrated in Example 2 below, the Griffin workflow can be leveraged to accurately distinguish these cell types from input sequence reads generated from ctDNA. Of course, it will be appreciated that the disclosure also encompasses embodiments of distinguishing determining the status of multiple features to precisely subtype the prostate cancer in the subject. For example, determining the cancer subtype comprises determining 2, 3, 4, or all of the following: whether the cancer is AR+ (ARPC) or AR-, whether the cancer is AR-low or ARPC, whether the cancer has a neuroendocrine prostate cancer (NEPC) phenotype signature or not, whether the cancer is AR-low or NEPC, whether the cancer is amphicrine or ARPC or NEPC, in any combination.

In another embodiment, the cancer is characterized as metastatic lung cancer. In a further embodiment, determining the subtype of the lung cancer comprises determining whether the cancer is small cell lung cancer (SCLC) or non-small cell lung cancer (NSCLC). If the lung cancer is NSCLC, in a further embodiment, the method further comprises determining whether the NSCLC is adenocarcinoma or squamous cell carcinoma.

As indicated above, the input sequence read data can be generated from a variety of platforms and with a variety of techniques, including whole genome analysis. In Example 3, the inventors established that whole genome analysis, however, is not required. Instead, the inventors designed and implemented a panel of genomic targets deemed to be relevant to the scientific inquiry (e.g., subtyping lung cancer cells).

Accordingly, in some embodiments, the lung cancer is further subtypes using sequence read data generated from a panel of genomic targets. In some embodiments, the panel of genomic targets comprises transcription factor binding sites (TFBSs) of one or more transcription factors associated with a designated subtype that is the subject of analysis, e.g., SCLC. For example, for subtyping SCLC, the one or more associated transcription factors comprise one or more of ASLC, NEUROD1, POU2F3, REST, and the like. In such embodiments, the method comprises determining the nucleosome occupancy of the TFBSs using any appropriate technique (e.g., CUT & RUN, and the like). The TFBSs can be identified by ChIP-seq data, or similar techniques known in the art. Candidate TFBSs can be retained in the panel if they are proximal to a transcription start site (TSS) of a gene associated with lung cancer, or the subtype of lung cancer that is of interest in the subtyping. In this regard, the term proximal can mean within a proximity that the TFBSs is functionally influential on the start of transcription at the TSS. In some instances, the functional influence or relationship can be established if the TSS is the closest TSS to the TFBS. In other embodiments, the panel of genomic targets comprise transcription start sites (TSSs) for one or more markers associated with lung cancer (or the specific subtype of lung cancer that is of interest). In such embodiments, the method comprises determining the nucleosome occupancy of the TSSs through known techniques.

The biological sample described herein can be any sample obtained from a subject that is likely to have cell free DNA. Illustrative, non-limiting examples encompassed by the disclosure include the sample is blood, plasma, or serum, which are particularly useful to assess cfDNA and ctDNA from a subject. In any embodiment of the foregoing aspects relating to detection or assessment of cancers in a subject, the methods can further comprise obtaining the biological sample from the subject. Additionally, for a subject that is determined to have cancer or a cancer subtype at any time, the method can further comprise prescribing appropriate treatment or actively treating the subject appropriately based on the determination of the cancer type or subtype according to accepted practice in the medical field for the determined cancer.

In any aspect described herein, the described method can be performed multiple times to provide multiple assessments. This can be useful to provide methods for monitoring the presence or evolution of cell types or subtypes from a source. For example, the methods can be performed from sequence read data obtained from biological samples obtained from a subject before and/or for time points at or after initial diagnosis of cancer.

As indicated above, the Griffin workflow is flexible and is not limited to a certain set of genomic regions of interest, nor to a specific type of sequence data for generating coverage profiles. Exemplary, non-limiting approaches for generating sequence read data include whole genome sequencing (for example depths between 0.05X coverage and 100X coverage) and chromatin accessibility assays. In some embodiments, the sequence read data is generated by, or regions of interests are identified using, techniques such as ATAC-seq, ChIP-seq, DNAse sensitivity assays, and the like, which are known in the art. In some embodiments, the sequence data is generated by CUT & RUN. See, e.g., WO 2019/060907, incorporated herein by reference in its entirety. For example, the CUT & RUN assay can incorporate use of one or more affinity reagents (e.g., antibodies or antibody fragments) that target post-translational modifications of H3K27ac, H3K4me1 and/or H3K27ac. In some embodiments, the method comprises affirmatively generating the sequence read data, using for example, any of the illustrative approaches described herein or other appropriate approaches known in the art.

As described above in the context of the lung cancer subtyping, the sequence read data can be produced from a panel of genomic targets. It will be understood that this targeted panel approach is applicable beyond Lung cancer subtyping to other types of cancers. Thus, in some embodiments of any of the methods described above, the sequence read data can comprise sequence read data generated from a panel of genomic targets. The panel of genomic targets can be designed and assembled according to the approach described in Example 3 in the context of lung cancer (see also FIG. 16). For example, the panel can comprise TFBSs of one or more transcription factors associated with a cancer type of interest. The transcription factors associated with a cancer type of interest can be readily identified from the art. The TFBSs relating to the designated transcription factor(s) can be determined by standard assays that establish binding sites in the genome, such as ChIP-seq data, and the like. Furthermore, candidate TFBSs can be further retained based on an assessment of association or proximity with transcription start sites (TSSs) of genes with transcription levels (on, off, high, low, etc.) associated with a relevant cancer or cancer subtype. In some embodiments, the panel of genomic targets comprise transcription start sites (TSSs) for one or more markers associated with the cancer type of interest. The panel can be constructed using the TFBSs and/or TSSs in any combination. Once established, directed sequencing reads are generated from the targets. In some embodiments, the nucleosome occupancy of the TFBSs and/or TSSs is determined. The sequence read data is the input into the computer-implemented Griffin method described above to facilitate the appropriate subtyping or other analysis.

As disclosed in Example 2, an analysis of fragment size variation using a coefficient of variation to assess standard deviation provided a metric that had a surprisingly strong correlation with gene expression/activity. This analysis can be integrated with or independent from performance of the Griffin workflow. Accordingly, in another aspect, the disclosure provides a computer-implemented method of enhancing sequence read data from cell-free DNA samples for cell type prediction. The method comprises:

receiving, by a computing system, sequence read data, wherein the sequence read data includes a plurality of fragment reads, and wherein each fragment read has a fragment length;

determining, by the computing system, a fragment size variability for at least one gene associated with a cell type; and

predicting, by the computing system, the cell type based on the fragment size variability for the at least one gene.

FIG. 23 is a flowchart that illustrates a non-limiting example embodiment of enhancing sequence read data from cell-free DNA samples for improved cell type prediction according to various aspects of the present disclosure.

At block 702, a computing system receives sequence read data, wherein the sequence read data includes a plurality of fragment reads, and wherein each fragment read has a fragment length.

At block 704, the computing system determines a fragment size variability for at least one gene associated with a cell type. In some embodiments, locations of genes whose mRNA expression and transcriptional activity are known to be associated with given cell types, such as the 47 genes illustrated in Fig. 12D that are known to be associated with prostate cancer, may be used.

In some embodiments, a coefficient of variation of the fragment size of fragments at locations associated with one or more genes may be determined and used as fragment size variability values. The coefficient of variation (CV) has been found to be particularly useful in distinguishing cell types based on fragment size variability when analyzing fragments at genes that are associated with the cell types. In particular, CV has been found to be less affected by the depth of sequencing coverage than other techniques (such as measurements of entropy).

At block 706, the computing system predicts the cell type based on the fragment size variability for at least one gene. In some embodiments, features may be generated based on the fragment size variability, and the features may be provided as input to a classifier model to determine whether the features represent a given cell type. In one non-limiting example, a ratio of the fragment size variability in a first cell type versus a second cell type may be used as a feature. The classifier model may be used to determine whether the calculated features for a given sample are more like features of a first cell type or a second cell type. Any suitable classifier model, including but not limited to logistic regression models, artificial neural networks, decision trees, support vector machines, and Bayesian networks, may be used.

One non-limiting example embodiment of the use of the method 700 is described in Example 2, where analysis of fragment size variability is used to distinguish prostate cancer cell types of androgen receptor pathway active prostate cancer (ARPC) varieties and neuroendocrine prostate cancer (NEPC) varieties.

Additional definitions

Unless specifically defined herein, all terms used herein have the same meaning as they would to one skilled in the art of the present invention. Practitioners are particularly directed to Sambrook J., et al. (eds.), *Molecular Cloning: A Laboratory Manual*, 3rd ed., Cold Spring Harbor Press, Plainsview, New York (2001); and Ausubel, F.M., et al. (eds.), *Current Protocols in Molecular Biology*, John Wiley & Sons, New York (2010); for definitions and terms of art.

The use of the term "or" in the claims is used to mean "and/or" unless explicitly indicated to refer to alternatives only or the alternatives are mutually exclusive, although the disclosure supports a definition that refers to only alternatives and "and/or."

Following long-standing patent law, the words "a" and "an," when used in conjunction with the word "comprising" in the claims or specification, denotes one or more, unless specifically noted.

Unless the context clearly requires otherwise, throughout the description and the claims, the words "comprise," "comprising," and the like, are to be construed in an inclusive sense as opposed to an exclusive or exhaustive sense; that is to indicate, in the sense of "including, but not limited to." Words using the singular or plural number also include the plural and singular number, respectively. Additionally, the words "herein," "above," and "below," and words of similar import, when used in this application, shall refer to this application as a whole and not to any particular portions of the application. The word "about" indicates a number within range of minor variation above or below the stated reference number. For example, "about" can refer to a number within a range of 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, or 1% above or below the indicated reference number.

The terms "subject," "individual," and "patient" are used interchangeably herein to refer to a mammal being assessed for treatment and/or being treated. In certain embodiments, the mammal is a human. The terms "subject," "individual," and "patient" encompass, without limitation, individuals having cancer. While subjects may be human, the term also encompasses other mammals, particularly those mammals useful as laboratory models for human disease, e.g., mouse, rat, dog, non-human primate, and the like.

The term "treating" and grammatical variants thereof may refer to any indicia of success in the treatment or amelioration or prevention of a disease or condition (e.g., a cancer, infectious disease, or autoimmune disease), including any objective or subjective parameter such as abatement; remission; diminishing of symptoms or making the disease condition more tolerable to the patient; slowing in the rate of degeneration or decline; or making the final point of degeneration less debilitating.

The treatment or amelioration of symptoms can be based on objective or subjective parameters; including the results of an examination by a physician. Accordingly, the term "treating" includes the administration of the compounds or agents of the present disclosure to prevent or delay, to alleviate, to improve clinical outcomes, to decrease occurrence of symptoms, to improve quality of life, to lengthen disease-free status, to stabilize, to prolong survival, to arrest or inhibit development of the symptoms or conditions associated with a disease or condition (e.g., a cancer), or any combination thereof. The term "therapeutic effect" refers to the reduction, elimination, or prevention of the disease or condition, symptoms of the disease or condition, or side effects of the disease or condition in the subject.

As used herein, the terms "nucleic acid" or "polynucleic acid" refer to a polymer of nucleotide monomer units or "residues", typically DNA or RNA. The nucleotide monomer subunits, or residues, of the nucleic acids each contain a nitrogenous base (i.e., nucleobase) a five-carbon sugar, and a phosphate group. The identity of each residue is typically indicated herein with reference to the identity of the nucleobase (or nitrogenous base) structure of each residue. Canonical nucleobases include adenine (A), guanine (G), thymine (T), uracil (U) (in RNA instead of thymine (T) residues) and cytosine (C). However, the nucleic acids of the present disclosure can include any modified nucleobase, nucleobase analogs, and/or non-canonical nucleobase, as are well-known in the art.

Disclosed are materials, compositions, and components that can be used for, can be used in conjunction with, can be used in preparation for, or are products of the disclosed methods and compositions. It is understood that, when combinations, subsets, interactions, groups, etc., of these materials are disclosed, each of various individual and collective combinations is specifically contemplated, even though specific reference to each and every single combination and permutation of these compounds may not be explicitly disclosed. This concept applies to all aspects of this disclosure including, but not limited to, steps in the described methods. Thus, specific elements of any foregoing embodiments can be combined or substituted for elements in other embodiments. For example, if there are a variety of additional steps that can be performed, it is understood that each of these additional steps can be performed with any specific method steps or combination of method steps of the disclosed methods, and that each such combination or subset of combinations is specifically contemplated and should be considered disclosed. Additionally, it is understood that the embodiments described herein can be implemented using any suitable material such as those described elsewhere herein or as known in the art.

Publications cited herein and the subject matter for which they are cited are hereby specifically incorporated by reference in their entireties.

## EXAMPLES

The following examples are set forth so as to provide those of ordinary skill in the art with a complete disclosure and description of how to make and use the present invention, and are not intended to limit the scope of what the inventors regard as their invention nor are they intended to represent that the experiments below are all or the only experiments performed.

## Example 1

This Example describes a study providing a proof-of-concept demonstration that sequence analysis applying an embodiment of the Griffin workflow disclosed herein enhances sequence signals with sufficient power and specificity to allow determination of breast cancer subtypes. Elements of this work are also described in Doebley, A.-L., et al. (2021). Griffin: Framework for clinical cancer subtyping from nucleosome profiling of cell-free DNA. MedRxiv 2021.08.31.21262867, incorporated herein by reference in its entirety.

### Introduction

Accurate cancer diagnosis and subtype classification are critical for guiding clinical care and precision oncology. Current approaches to determine tumor subtype require a tissue biopsy, which is often difficult to obtain from patients with metastatic cancer. Therefore, at the time of recurrence or metastatic cancer diagnosis, treatment options may often be informed by clinical diagnostics from the primary tumor. However, molecular changes in the tumor can emerge during metastatic progression and in the context of therapeutic resistance. Moreover, surveying molecular changes is challenging because repeated biopsies are problematic and not routine in clinical practice for solid tumors.

Cell-free DNA (cfDNA) is DNA released into circulation by cells during apoptosis and necrosis. In patients with cancer, a portion of this cfDNA is released from tumor cells, called circulating tumor DNA (ctDNA). The analysis of ctDNA can address the challenges in tissue accessibility and has demonstrated great potential for clinical utility. Much of the current research and clinical efforts have focused on the detection of genetic alterations in ctDNA. Shallow coverage sequencing of cfDNA, including ultra-low pass whole genome sequencing (ULP-WGS, 0.1x), provides a cost-effective and scalable solution for estimating the tumor fraction (fraction of the cfDNA that is tumor derived) from the analysis of genomic copy number alterations. Sequencing analysis of genomic alterations from ctDNA have helped to distinguish molecular subsets of tumors. However, these genomic alterations, including somatic mutations, may not always fully explain treatment failure or identify therapeutic targets, exemplifying a major limitation of cancer precision medicine.

Tumor subtypes are often characterized by distinct transcriptional regulation, which can change during treatment resistance, leading to different clinical tumor phenotypes. For example, prostate and lung cancers may undergo trans-differentiation from adenocarcinoma to small-cell neuroendocrine phenotypes. For metastatic breast cancer (MBC), treatment is guided based on clinical subtypes determined by the expression of the estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2), often in the primary tumor; endocrine therapies are prescribed to patients with ER-positive (ER+) or PR-positive (PR+) carcinomas while patients with HER2 positive tumors are prescribed anti-HER2 drugs. Patients with tumors absent for expression of all three receptors have triple negative breast cancer (TNBC) and receive chemotherapy. However, receptor conversions during primary and metastatic disease progression have been frequently observed, including ~20% of patient tumors switching from ER+ to ER-negative (ER-) subtypes. Furthermore, similar to the presence of intra-tumor genomic heterogeneity in breast cancer, mixtures of clinical subtypes may also co-exist across or within metastatic lesions in the same patient, presenting major clinical challenges. Therefore, accurate subtype classification and identification of transcriptional patterns underlying emergent clinical phenotype during therapy has critical implications for studying mechanisms of resistance and informing treatmentdecisions.

Recent studies have shown that the computational analysis of cfDNA fragmentation patterns from genome sequencing data can reveal the occupancy of nucleosomes in cells-of-origin. When DNA is released into the peripheral blood following cell death, they are protected from degradation by nucleosomes. At accessible genomic locations, such as at actively bound transcription factor binding sites (TFBSs) and open chromatin regions, nucleosomes are positioned in an organized manner that allows access for DNA binding proteins (FIG. 7A). This nucleosome organization results in a loss of sequencing coverage, reflecting DNA degradation at the unprotected binding site with peaks of coverage at the surrounding protected locations.

Applications of nucleosome profiling from cfDNA have been demonstrated for cancer detection and tumor tissue-of-origin prediction, including the analysis of shorter cfDNA fragments which tend to be enriched from tumor cells. While tumor subtyping from cfDNA has been explored in prostate cancer by analyzing TFBS locations, it is believed that there have not been demonstrations of subtype classification from cfDNA in other cancers. Specifically, predicting histological subtypes in breast cancer has not been shown from cfDNA. Furthermore, current cfDNA nucleosome profiling approaches have not been optimized for ULP-WGS data. Studying the clinical phenotype of tumors from ctDNA remains challenging due to lack of robust computational methods but has obvious potential clinical benefits for guiding treatment decisions in patients with metastatic cancer.

In this present study, a computational framework called Griffin was developed to classify tumor subtypes from nucleosome profiling of cfDNA. Griffin overcomes current analytical challenges to profiles the nucleosome accessibility and transcriptional regulation from the analysis of standard cfDNA genome sequencing, including ULP-WGS (0.1x) coverage. Griffin employs a novel GC correction procedure that is specific for DNA fragment sizes and therefore unique for cfDNA sequencing data. Griffin was applied to perform cancer detection and tumor tissue-of-origin analysis with high performance. Then, the first application of breast cancer ER subtyping from cfDNA was demonstrated, showing strong classification accuracy and insights into tumor heterogeneity and prognosis, all achieved from analysis of ULP-WGS data. Overall, Griffin is a generalizable framework that can detect molecular changes in transcriptional regulation and chromatin accessibility from cfDNA and possibly direct personalized treatment to improve patient outcomes.

Results

*Griffin framework for nucleosome profiling to predict tumorphenotype*

Griffin was developed as an analysis framework with a GC correction procedure to accurately profile nucleosome occupancy from cfDNA. Griffin processes fragment coverage to distinguish accessible and inaccessible features of nucleosome protection (FIG. 7A). Griffin is designed to be applied to whole genome sequencing (WGS) data of cfDNA from patients with cancer to quantify nucleosome protection around sites of interest and is optimized to work for ULP-WGS data (FIG. 7B). Sites of interest can be selected from various chromatin-based assays, such as from assay for transposase-accessible chromatin using sequencing (ATAC-seq) and are tailored to address specific problems including cancer detection and tumorsubtyping.

The analysis workflow begins with computing the genome-wide fragment-based GC bias for each sample. Then, for the region at each site of interest, the fragment midpoint coverage is computed and reweighted to remove GC biases (Methods). Midpoint coverage rather than full fragment coverage is used because it produces higher amplitude nucleosome protection signals (not shown). Next, a composite coverage profile is computed as the mean of the GC- corrected coverage across the set of sites specific for a tissue type, tumor type, transcription factor (TF), or any phenotypic comparison of interest. By examining these coverage profiles around known cancer-specific and blood-specific TFs, three quantitative features were identified that distinguish a site as accessible and inaccessible: (a) the coverage in the window between - /+ 30 bp ('central coverage'), where lower values represent increased accessibility, (b) the coverage in a window between -/+ 1000 bp ('mean coverage'), and (c) the overall nucleosome peak amplitude calculated using Fast Fourier transform ('amplitude'). These features can be used to quantify transcription factor activity or chromatin accessibility and be used as features for detection of cancer, tumor subtyping, or studying other phenotypes of interest.

*Griffin reduces GC biases enabling detection of tissue specificaccessibility*

A novel aspect of Griffin is the implementation of a fragment-based GC bias correction. At open chromatin regions, especially at TFBS, GC-content is non-uniform, which leads to GC-related coverage biases (FIG 8A) (Wang, J. et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. 22, 1798–1812 (2012)). GC bias varies between samples and between different fragment lengths within a sample (Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Research 40, e72–e72 (2012)) (FIG. 8B), which can have a major impact on nucleosome accessibility prediction (FIG. 8C). To correct for this GC bias, for each sample and each fragment length, Griffin computes the global estimated mean fragment coverage ("expected") using a fragment length position model (Benjamini, Y. & Speed, T. P. Nucleic Acids Research 40, e72–e72 (2012)) (Methods, FIG. 8B). Then, when calculating coverage profiles around sites of interest, each fragment is assigned a weight based on the global expected coverage for its length and GC bias. This correction eliminates unexpected increases (or decreases) in coverage at binding sites, removing technical biases to enhance the tissue-associated accessibility signals when analyzing WGS (9-25x, FIG. 8C) cancer patient cfDNA and ULP-WGS (0.1-0.3x, FIG. 8D).

To test the performance of nucleosome profiling following Griffin GC-bias correction, the estimated TFBS accessibility was compared with the amount of tumor-derived DNA (i.e. tumor fraction) predicted by ichorCNA for ULP-WGS data from 191 MBC cfDNA samples with ≥ 0.1 tumor fraction (Adalsteinsson, V. A. et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nature Communications 8, (2017)). The tumor fraction was expected to be negatively corrected with the central coverage around tumor-specific sites, and positively correlated for blood-specific sites. For a blood specific TF, LYL1, it was observed that the central coverage at TFBSs was positively correlated with tumor fraction before GC correction (Pearson's r=0.41) as expected, but this correlation was much stronger after GC correction (Pearson's r=0.63, FIG.8E). For a tumor-specific TF, GRHL2, a negative correlation was observed between the central coverage and tumor fraction, as expected (Pearson's r=-0.62, not shown). The mean coverage and amplitude features are also correlated to tumor fraction but appeared to be less influenced by GC bias (not shown). Similar correlations between nucleosome profile features and tumor fraction following GC correction were also observed for blood and cancer specific DNase I hypersensitivity sites (DHSs) (not shown).

To quantify how GC correction reduces signal variability between samples, the central coverage in the 191 MBC cfDNA ULP-WGS samples was examined for 377 TFs in the Gene Transcription Regulation Database (GTRD) (Ulz, P. et al. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nature Communications 10, 4666 (2019); Yevshin, I., et al. GTRD: A database on gene transcription regulation - 2019 update. Nucleic Acids Research 47, D100–D105 (2019)). For each factor, the variability between the central coverage and tumor fraction using the root mean squared error (RMSE) from a linear regression fit was compared before and after GC correction. For LYL1, the RMSE decreased (0.062 to 0.046), indicating less inter-sample variation in the data after GC correction (FIG. 8E). Similarly, for 351 (93.1%) TFs, the RMSE was decreased after GC correction, indicating reduced inter-sample variability after accounting for the correlation between tumor fraction and central coverage (two-sided Wilcoxon signed rank test p = $1.0\text{x}10^{-58}$, test statistic = 1421, FIG. 8F). Additionally, the central coverage for the 377 TFs was examined in a cohort of 215 healthy donors (Cristiano, S. et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature 570, 385–389 (2019)) before and after GC correction. Because healthy donor samples have no tumor content, the mean absolute deviation (MAD) was evaluated for each TF to compare inter-sample variability. It was found that the MAD decreased after GC correction for 365 (96.8%) TFs (two-sided Wilcoxon signed rank test p = $6.28\text{x}10^{-62}$, test-statistic = 466, FIG. 8G), indicating lower inter-sample variability for nearly all TFs. Altogether, these results demonstrate that the GC correction in the Griffin framework reduces the variability in chromatin accessibility signals due to GC biases between samples and allows for improved detection of tissue specific accessibility in ULP-WGS data.

*Griffin analysis at TFBS enables accurate cancer detection*

To determine if Griffin can perform cancer detection, a published WGS (1-2X) dataset of cfDNA samples from healthy donors (n = 215) and cancer patients (n = 208) was analyzed (Cristiano, S. et al. Nature 570, 385–389 (2019)). We generated nucleosome profiles around TFBSs for the 377 TFs using nucleosome sized (100-200bp) fragments and extracted three features from each profile (central coverage, mean coverage, and amplitude) for a total of 1131 features. Using logistic regression, a high performance for predicting the presence of cancer was achieved with an area under the receiver operating curve (AUC) of 0.94 (FIG.9B). The highest performance was observed for stage IV cancers (AUC=0.99), with a lower performance for stage I cancers (AUC=0.93). The performance was likely reflective of the higher tumor fractions observed in late-stage cancer relative to early-stage cancer. Higher performance was observed for samples with tumor fraction ≥ 0.05 (AUC 0.99) than samples with undetectable tumor (0 tumor fraction, AUC=0.90). Somewhat lower performance was observed with Griffin analysis around DNase I Hypersensitivity Sites (DHS) (AUC=0.83).

To test the ability to detect cancer at ULP-WGS coverage (0.1x), Griffin was applied to the same cfDNA data downsampled to 0.1x coverage and achieved a performance with AUC of 0.89 (FIG.9B). Next, because fragments <150bp are enriched for tumor derived DNA (Cristiano, S. et al. Nature 570, 385–389 (2019)), it was tested whether using only shorter fragments might improve the ability to detect cancer in this framework. Griffin was applied to analyze only 35-150bp fragments at the same TFBSs and observed a decreased performance (AUC=0.91, not shown). Finally, the results were compared with the method by Ulz et al. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nature Communications 10, 4666 (2019), which analyzed cfDNA fragments of all lengths at TFBSs and achieved an AUC of 0.82 for 1-2x data and 0.55 for downsampled data (not shown). Griffin using nucleosome-sized or short fragments and ULP-WGS coverage had higher detection performance. This demonstrates that Griffin can detect cancer accurately using various sites from chromatin-based assays and cost-effective ULP-WGS of cfDNA.

*Griffin enables accurate prediction of breast cancer subtypes from ultra-low pass WGS*

Breast cancer tumor classification relies on accurate clinical determination of hormone receptor status primarily by immunohistochemistry (IHC) to quantify the expression of ER, but no ctDNA approach exists for this application. Thus, a goal was to determine whether Griffin can be used to predict ER subtype status from ULP-WGS (0.1x) of cfDNA from MBC patients. 254 samples (Adalsteinsson, V. A. et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nature Communications 8, (2017); Stover, D. G. et al. Association of Cell-Free DNA Tumor Fraction and Somatic Copy Number Alterations With Survival in Metastatic Triple-Negative Breast Cancer. Journal of Clinical Oncology JCO.2017.76.003 (2018)) with tumor fraction greater than 0.05 from 139 patients were analyzed. First, the Griffin profiles were inspected at TFBSs for key factors, including ESR1, FOXA1, and GATA3, which are known to be associated with ER positive tumors (Albergaria, A. et al. Expression of FOXA1 and GATA-3 in breast cancer: the prognostic significance in hormone receptor-negative tumours. Breast Cancer Research 11, R40 (2009)). It was observed that these TFBSs were more accessible in cfDNA samples from patients with ER+ metastases compared to ER-; central coverage was significantly lower in ER+ samples after accounting for tumor fraction (ANCOVA q-value for ER status < $3.38x10^{-2}$, not shown). To predict ER status, a logistic regression classifier was initially built using features from the Griffin profiles for all 377 TFs and an accuracy of 0.71 (AUC of 0.79) was achieved (not shown). TFBSs features computed by the Ulz method were also used for ER subtyping and an accuracy of 0.53 (AUC=0.55) was observed (not shown) likely because it was not designed for ULP-WGS data.

Next, a more tailored site selection approach was used by analyzing regions of differential chromatin accessibility. Using ATAC-seq data generated from 44 ER+ and 15 ER- primary breast tumors by The Cancer Genome Atlas (TCGA) (Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. Science 362, eaav1898 (2018)), open chromatin sites that were specific to each ER subtype were identified (Methods, FIG. 10A). ER+ specific sites (n=28,170) were enriched for the TFBSs of ESR1, PGR, FOXA1 and GATA3, and ER- specific sites (n=41,712) were enriched for the TFBSs of STAT3 and NFKB1 (not shown). Differences were observed in coverage profiles between ER subtype-specific sites that were shared and not shared with accessible chromatin in hematopoietic cells (Satpathy, A. T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nature Biotechnology 37, 925–936 (2019)) and, thus, the were analyzed separately (FIG. 10B).

Griffin was applied to profile nucleosome accessibility at these four sets of ER subtype-specific accessible chromatin sites, extracting a total of 12 features (FIG. 10B). A logistic regression classifier was built to predict ER subtype from these chromatin accessibility features and achieved an overall accuracy of 0.81 (AUC=0.89, n=139) (Methods, FIG. 10D). The performance was higher for samples with high tumor fraction (accuracy 0.86, AUC=0.92, n=101, tumor fraction ≥ 0.1) compared to those with lower tumor fraction (accuracy 0.69, AUC=0.75, n=38, tumor fraction 0.05 to 0.1) (FIG. 10D). Repeating the analysis using only short fragments (35-150bp) did not improve the performance (accuracy 0.73, AUC=0.81), likely due to further reduced fragment coverage (not shown). These results illustrate the utility of using chromatin accessibility for cancer subtyping from ULP-WGS data and showcase the first application of ER status prediction in breast cancer from cfDNA. We validated this finding by examining cfDNA samples from two other studies of patients with breast cancer and one new dataset (see Ahuno ST, et al. Ghana Breast Health Study Team. Circulating tumor DNA is readily detectable among Ghanaian breast cancer patients supporting non-invasive cancer genomic studies in Africa. NPJ Precis Oncol. 5(1):83 (2021) and Zivanovic Bujak, A., et al. Circulating tumour DNA in metastatic breast cancer to guide clinical trial enrolment and precision oncology: A cohort study. PLoS medicine 17.10 (2020): e1003363) and using the model trained on the original MBC dataset, we were able to predict ER status with 0.92 accuracy (0.96 AUC) in all samples with >0.05 tumor fraction. Looking only at samples with >0.1 tumor fraction, the accuracy was 0.96 and the AUC was 0.98. This analysis further supports that Griffin can perform accurate ER status prediction in independent datasets.

*Analysis of ER status from cfDNA suggests tumor subtype heterogeneity*

To further investigate the ER predictions, the classification results were inspected to look for patterns in the tumor fraction and primary tumor ER status for samples with incorrect predictions (FIG. 10C). We observed that the many of the incorrect predictions were in samples with ER loss (ER+ primary and ER- metastasis) (FIG. 10F) and that the number of ER+ predictions for these patients (5 ER+ predicted out of 9 total ER loss patients) were significantly different than the number of ER+ predictions in both ER- patients with ER- primary (two-sided Fisher's exact test p = 3.7X10$^{-4}$) and ER+ patients who retained ER+ in both primary and metastasis (4.3x10$^{-2}$, two-sided Fisher's exact test p = 0.0183.7X10$^{-4}$, FIG. 10F). However, despite incorrectly predicting that many of the ER loss patients remained ER+, we found that an ROC analysis of metastatic ER status predictions among patients with ER+ primary and tumor fraction >0.1 resulted in an AUC of 0.74, suggesting that Griffin has a reasonable ability to detect ER loss among patients with ER+ primary (FIG. 10G). However, the overall lower performance among ER loss patients compared to patients with unchanged ER status, suggesting that there may be residual ER+ tumor features in the ER loss patients or that Griffin analysis may be capturing a heterogeneous mixture of ER subtypes from ctDNA.

To further assess whether this observation may be due to tumor heterogeneity, we examined the other metastatic biopsies taken in patients with ER loss. We found that in many of these cases, the patients had additional ER+ metastatic biopsies after an initial ER- diagnosis. Two particularly interesting cases are shown in FIG. 10H.Patient MBC_1413 was initially diagnosed with ER- metastatic disease in a pleural fluid biopsy, however they later had a second metastatic biopsy of a liver metastasis which showed ER expression in 5% of cells. The first cell-free DNA blood draw was taken shortly after this biopsy and interestingly was predicted to be ER+ in agreement with the ER low status of the liver biopsy. Later, a third biopsy was taken from pleural fluid and once again showed ER- disease. Shortly after this biopsy, a cfDNA blood draw was taken and this blood draw was predicted to be ER-. In a second patient, MBC_1099, the first two metastatic biopsies (bone and liver) showed ER- disease. However, when cfDNA was drawn a few months later, the patient's subtype was predicted to be ER+ for two timepoints. Interestingly, when another liver biopsy was taken between the two cfDNA blood draws, this biopsy showed 5% ER+ cells, potentially explaining the ER+ prediction from the cfDNA. These results suggest that in some cases where Griffin failed to detect an ER loss, the predictions may be detecting true ER subtype heterogeneity and suggest that Griffin could be used to monitor subtype dynamics over the course of therapy.

## Discussion

In this study, the development of Griffin, a new framework and analysis tool for studying transcriptional regulation and tumor phenotypes, is described. Griffin uses a novel cfDNA fragment length-specific normalization of GC-content biases that obscure chromatin accessibility information. It is demonstrated that Griffin can be used to detect cancer from low pass WGS with high accuracy. Additionally, an approach was developed to perform ER subtyping in breast cancer from ULP-WGS, which is the first time that ER phenotype prediction has been shown from ctDNA.

Griffin is versatile and can be used for various applications in cancer. This disclosure highlights cancer detection, tissue-of-origin, and tumor subtype use-cases. However, Griffin can also be used for any biological comparison where transcriptional regulation and chromatin accessibility differences can be delineated. The applications described here use TFBSs from chromatin immunoprecipitation sequencing (ChIP-seq) and accessible chromatin sites from ATAC-seq. However, Griffin differs from existing methods due to its ability to analyze custom sites of interest that are specific to any biological context. These sites may be obtained from external sources and different assays, such as ChIP-seq, DNase I hypersensitivity, ATAC-seq or cleavage under targets and release using nuclease (CUT&RUN). As additional epigenetic data are collected by the cancer research community, including from single-cell experiments (Wu, S. J. et al. Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. Nat Biotechnol 39, 819–824 (2021); Pierce, S. E., Granja, J. M. & Greenleaf, W. J. High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. Nat Commun 12, 2969 (2021)), Griffin will be integral for advancing tumor phenotype studies from liquid biopsies.

Griffin is optimized for the analysis of ULP-WGS (0.1x) of cfDNA, while other nucleosome profiling methods have focused on deeper coverage sequencing. Griffin takes advantage of analyzing the breadth of sites as opposed to individual loci, which was inspired by a similar strategy used by Ulz, P. et al. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nature Communications 10, 4666 (2019). It is demonstrated that Griffin has better performance for both detecting cancer and predicting ER status from ULP-WGS data when compared to the Ulz method, because of its novel bias correction and versatility to analyze any set of genomic regions. However, Griffin is not limited to low coverage data. Increased cfDNA sequencing coverage can allow for analysis of specific gene promoters and cis-regulatory elements and may be able to inform gene expression (Ulz, P. et al. Inferring expressed genes by whole-genome sequencing of plasma DNA. Nature Genetics 48, 1273–1278 (2016)). While recent studies show the promise of cfDNA methylation and cfRNA analysis for tumor phenotype analysis and cancer detection (Beltran, H. et al. Circulating tumor DNA profile recognizes transformation to castration- resistant neuroendocrine prostate cancer. J Clin Invest 130, 1653–1668 (2020); Wu, A. et al. Genome-wide plasma DNA methylation features of metastatic prostate cancer. J Clin Invest 130, 1991–2000 (2020); Shen, S. Y. et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature 563, 579–583 (2018); Liu, M. C. et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Annals of Oncology 31, 745–759 (2020); Larson, M. H. et al. A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection. Nature Communications 12, 2357 (2021); Kang, S. et al. CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. Genome Biology 18, 53 (2017); Chan, K. C. A. et al. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. Proceedings of the National Academy of Sciences 110, 18761–18768 (2013)), these analytes may be challenging to isolate from clinical specimens or require specialized assays. Griffin provides a cost-effective and scalable method requiring only standard low coverage WGS of cfDNA, which can be more rapidly incorporated into existing platforms to predict clinical cancer phenotypes.

A limitation of the binary ER classification (ER+ or ER-) is the decreased accuracy for samples with lower tumor fraction (0.05 to 0.1); however, patients with cfDNA tumor fraction $\geq 10\%$ have poorer prognosis (Stover, D. G. et al. Association of Cell-Free DNA Tumor Fraction and Somatic Copy Number Alterations With Survival in Metastatic Triple-Negative Breast Cancer. JCO 36, 543–553 (2018)) and would benefit more from tumor monitoring. It may be possible to improve performance of ER subtyping for lower tumor fraction samples with additional sequencing depth or joint analysis of multiple cfDNA timepoints from the same patient.

The application of Griffin to predict ER status from cfDNA of MBC patients led to interesting insights into tumor heterogeneity and potential explanations for misclassified predictions. Intriguingly, it was noted that for the patients with ER- tumors by IHC, ER+ predictions were significantly enriched when the primary tumor was ER+. Two patients with ER loss and ER+ predictions had metastasis of both subtypes. Importantly, while this subtype heterogeneity and switching would typically not be captured from a single metastatic biopsy, these results demonstrate the possibility of using ER probability to monitoring subtype heterogeneity over time during therapy using ctDNA.

The breast cancer subtyping was focused on ER prediction because its status has important utility in predicting likely benefit to endocrine therapy (Group (EBCTCG), E. B. C. T. C. Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. The Lancet 378, 771–784 (2011)). While PR expression is also determined in the clinic and ER-/PR+ tumors are considered hormone receptor positive, these are rare, not reproducible or less useful for prognosis (Hefti, M. M. et al. Estrogen receptor negative/progesterone receptor positive breast cancer is not a reproducible subtype. Breast Cancer Research 15, R68 (2013)). In the cohort, only 2 of 139 (1.4%) patients were ER-/PR+. HER2 overexpression is important relevant for prognosis and determining treatment such as trastuzumab (Slamon, D. J. et al. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science 235, 177–182 (1987)). However, an insufficient number of open chromatin sites were identified that were specific for distinguishing HER2 status. Since ERBB2 (encodes the HER2 protein) is amplified in ~20% breast cancers, one can instead assess ERBB2 copy number amplification from ctDNA genomic analysis (Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486, 346–352 (2012)). Alternatively, a model to predict PAM50 status could be useful as this may be a better indicator of prognosis than ER/PR/HER2 IHC alone (Nielsen, T. O. et al. A Comparison of PAM50 Intrinsic Subtyping with Immunohistochemistry and Clinical Prognostic Factors in Tamoxifen-Treated Estrogen Receptor–Positive Breast Cancer. Clinical Cancer Research 16, 5222–5232 (2010)).

The Griffin framework is a unique advance on our previous method to analyze genomic alterations and estimate tumor fraction from ULP-WGS of cfDNA (Adalsteinsson, V. A. et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nature Communications 8, (2017)). Together, these methods form a suite of tools to establish a new paradigm to study both tumor genotype and phenotype from ULP-WGS of cfDNA. Griffin has the potential to reveal clinically relevant tumor phenotypes, which will support the study of therapeutic resistance, inform treatment decisions, and accelerate applications in cancer precision medicine.

Methods

*Griffin: GC bias calculation*

GC content influences the efficiency of amplification and sequencing leading to different expected coverages (coverage bias) for fragments with different GC contents and fragment lengths. This is called GC bias and is unique to each sample. We calculated the GC bias of each bam file using an implementation of the method developed by Benjamini and Speed 2012 (Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Research 40, e72–e72 (2012)) which was previously implemented in deepTools (Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research 44, W160–W165 (2016)). However, unlike the deepTools implementation, which assumes that all fragments have the same length, we used the 'fragment length model' which calculates a separate GC bias curve for each fragment length. This is helpful for cfDNA where different samples may have different fragment size distributions and different fragment lengths have biological significance. Prior to performing GC bias calculation, we identified all mappable regions of the genome using the Umap multi-read mappability track for 50bp reads downloaded from UCSC genome browser (Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bismap: quantifying genome and methylome mappability. Nucleic Acids Research 46, e120–e120 (2018)) (hgdownload.soe.ucsc.edu/gbdb/hg38/hoffmanMappability/k50.Umap.MultiTrackMappability.bw). We used pybedtools (Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics 27, 3423–3424 (2011)) to find the mappable regions (defined as mappability score = 1) and further excluded regions with known mapping problems including the encode unified exclusion list (encodeproject.org/files/ENCFF356LFX/), centromeres, fix patches, and alternative haplotypes for hg38 downloaded from UCSC table browser (genome.ucsc.edu/cgi-bin/hgTables). We then examined all remaining regions of the genome and, for each fragment length, counted the observed GC content of every possible fragment overlapping those positions. The observed frequencies of each GC content for each fragment length are the 'genome GC frequencies'. We then developed the 'griffin GC bias' pipeline to compute the GC bias in a given bam file. The pipeline takes a bam file, bedGraph file of valid (mappable, non-excluded) regions, and genome GC frequencies for those regions. For each given sample, we fetched all reads aligning to the valid regions on autosomes using pysam (github.com/pysam-developers/pysam) (Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009)). We counted the number of observed reads for each length and GC content, excluding reads with low mapping quality (<20), duplicates, unpaired reads, and reads that failed quality control. These read counts are the 'GC counts' for that sample. We then divided the GC counts for a sample by the GC frequencies for the genome to obtain the GC bias for that bam file and normalized the mean GC bias for each fragment length to 1, resulting in a GC bias value for every combination of fragment size and GC content (except those combinations that are never observed in the genome). We then smoothed the GC bias curves. For each fragment size we took all GC bias values for fragments of a similar length (+/- 10 bp). We sorted these values by the GC content of the fragment to create a vector of GC bias values for similar sized fragments. We then smoothed this vector by taking the median of k nearest neighbors (where k = 5% of the vector length or 50, whichever is greater) and repeated for each possible fragment length. We then normalized to a mean GC bias of 1 for each possible fragment length (excluding GC contents that are never observed) to generate a smoothed GC bias value for every possible fragment length and GC content observed in the genome.

*Griffin: Nucleosome profiling*

We designed the griffin nucleosome profiling pipeline to perform nucleosome profiling around sites of interest. This pipeline takes a bam file and site list, and assorted other parameters described below. For a given bam file and site list, we fetched all reads in a window (-5000 to +5000bp) around each site using pysam (excluding those that failed quality control measures). We then filtered read pairs by fragment length and selected those in a range of fragment lengths (100-200 bp unless otherwise specified). For each read pair, we determined the GC bias for the fragment and assigned a weight of $\frac{1}{GC\ bias}$ to that fragment and identified the location of the fragment midpoint. We split the site into 15bp bins and summed the weighted fragment midpoints in each bin to get a GC corrected midpoint coverage profile (see Fig. 1b for a schematic). Next, we excluded bins that overlapped regions with known mapping problems (Griffin: described in GC bias calculation) and bins with at least one unmappable position. We also identified bins with extremely high coverage (10 standard deviations above the mean) and removed these bins. We repeated this for every site on the site list and took the mean of all sites (ignoring excluded bins within those sites) to generate the coverage profile for that site list. We then smoothed the coverage profiles using a Savitzky-Golay filter with window length 165bp and polynomial order of 3. Finally, to make samples with different depths comparable, we normalized the coverage profile to a mean coverage of 1 and retained the central region (+/- 1000 bp) for further analysis.

*Griffin: Nucleosome profile feature quantification*

To quantify coverage profiles, we extracted 3 features from each coverage profile. First, we calculated the 'mean coverage' value +/- 1000 bp from the site. Second, we calculated the coverage value at the site (+/- 30bp). And third, we calculated the amplitude of the nucleosome peaks surrounding the site by using a Fast Fourier Transform (as implemented in Numpy (Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020)) on the window +/-960 bp from the site and taking the amplitude of the 10th frequency term. This window and frequency were chosen due to the observed nucleosome peak spacing at an active site (190bp) which results in approximately 10 peaks in the window +/-960bp.

*Early-stage cancer and healthy donor cfDNA samples – DELFI dataset*

Whole genome sequencing (WGS) cfDNA from patients with various types of early stage cancer and healthy donors were obtained from an existing dataset published in Cristiano, S. et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature 570, 385–389 (2019). Bam files were downloaded from EGA (dataset ID: EGAD00001005339). This data consisted of 1-2x low pass whole genome sequencing from 100bp paired end Illumina sequencing reads. For our analyses, we used a subset of samples with 1-2X WGS of cfDNA from 208 cancer patients with no previous treatment and 215 healthy donors. These are the samples used for the cancer detection analysis in Cristiano et al. cfDNA tumor fraction was estimated using ichorCNA (Adalsteinsson, V. A. et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nature Communications 8, (2017) ). An hg38 panel of normal (PoN) with a 1mb bin size was created using all 215 healthy donors in the dataset. ichorCNA was then run on all cancer and healthy samples to estimate tumor fraction. ichorCNA_fracReadsInChrYForMale was set to 0.001. Defaults were used for all other settings.

*Early-stage lung cancer and healthy donor cfDNA samples – LUCAS dataset*

Whole genome sequencing (WGS) cfDNA from a prospective study of patients with lung cancer and without cancer were obtained from an existing dataset published by Mathios and colleagues (Mathios, D. et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. Nat Commun 12, 5060 (2021)). Bam files were downloaded from EGA (dataset ID: EGAD00001007796). This data consisted of 1-2x low pass whole genome sequencing from 100bp paired end Illumina sequencing reads. For our analyses, we used the subset of samples described in the paper as the 'LUCAS' cohort and a second subset of samples described as the validation cohort. The LUCAS cohort included 158 patients who had no history of cancer and no future cancer diagnosis and 129 patients who were diagnosed with lung cancer within days of blood draw (0-44 days). The validation cohort included 46 patients with cancer and 385 patients without cancer. All samples were realigned to hg38 as described below in sequence data processing. Tumor fraction was determined using ichorCNA as described above with a panel of normals constructed from 54 separate non-cancer samples from this same study.

*Metastatic breast cancer (MBC) and healthy donor cfDNA samples*

WGS of cfDNA from patients with metastatic breast cancer (MBC) and healthy donors were obtained from an existing dataset (Adalsteinsson, V. A. et al. Nature Communications 8, (2017)). Bam files were downloaded from dbGaP (accession code: phs001417.v1.p1). This data consisted of ~0.1x ultra-low pass whole genome sequencing (ULP-WGS) from 100bp paired end Illumina sequencing reads. For our analyses, we used a subset of 254 samples with >0.1X coverage WGS, >0.05X tumor fraction and known estrogen receptor (ER) status. Of these 254 samples 133 were ER positive (from 74 unique patients) and 121 were ER negative (from 65 unique patients). Coverage and tumor fraction metrics were obtained from the supplemental data in the publication (Adalsteinsson, V. A. et al. Nature Communications 8, (2017)).. Primary and metastatic ER status was determined by immunohistochemistry and abstracted from medical records. Additionally, we used deep (9-25X) WGS from two MBC patients (MBC_315 and MBC_288) from the same source and deep (17-20X) WGS from two healthy donors (HD45 and HD46) from the same source for designing and demonstrating the pipeline.

For training and assessing the ER status classifier we labeled each sample as ER+ or ER- using information about the ER status from medical records. If metastatic ER status was known, the sample was labeled according to this status. If metastatic ER status was not known, the sample was labeled according to the primary tumor ER status (20 samples from 11 patients). ER low samples (11 samples from 6 patients) were labeled ER positive for the purpose of the binary classifier. For three patients (MBC_1405, MBC_1406, MBC_1408), we had information about multiple metastatic biopsies with different ER statuses. In these cases, we used the last biopsy taken for the purpose of the binary ER status classifier.

*Human Subjects*

WGS of cfDNA samples from patients with MBC were obtained from an existing study as described above (Adalsteinsson, V. A. et al. Nature Communications 8, (2017)). Additional information, including primary ER status, metastatic ER status, and survival time, was abstracted from the medical records. Use of this data was approved by an institutional review board (Dana-Farber Cancer Institute IRB protocol identifiers 05-246, 09-204, 12-431 [NCT01738438; Closure effective date 6/30/2014]).

*Sequence data processing*

All sequencing data used in this study was realigned to the hg38 version of the human genome (downloaded from hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz). Bam files were unmapped from their previous alignment using Picard SamToFastq (Picard Toolkit. (Broad Institute, 2021)). They were then realigned to the human reference genome according to GATK best practices (DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43, 491–498 (2011)) using the following procedure. Fastq files were realigned using BWA-MEM (Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 00, 1–3 (2013)). Files were then sorted with samtools (Danecek, P. et al. Twelve years of SAMtools and BCFtools. GigaScience 10, 1–4 (2021)), duplicates were marked with Picard, and base recalibration was performed with GATK, using known polymorphisms downloaded from the following locations: console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0/Mills_and_1000G_gold_standard.indels.hg38.vcf.gz and ftp.ncbi.nih.gov/snp/organisms/human_9606_b151_GRCh38p7/VCF/GATK/All_20180418.vcf.gz.

*Transcription factor binding site (TFBS) selection*

Transcription factor binding sites (TFBSs) were downloaded from the GTRD database (Yevshin, I., GTRD: A database on gene transcription regulation - 2019 update. Nucleic Acids Research 47, D100–D105 (2019)). This database contains a compilation of ChIP seq data from various sources. For our analyses, we used the meta clusters data (version 19.10, downloaded from gtrd.biouml.org/downloads/19.10/chip-seq/Homo%20sapiens_meta_clusters.interval.gz). This contains meta peaks observed in one or more ChIP seq experiments. The GTRD database contains some ChIP seq experiments for targets that are not transcription factors (TFs). These were excluded by comparing against a list of TFs with known binding sites in the CIS-BP database (Weirauch, M. T. et al. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell 158, 1431–1443 (2014)) (v2.00 downloaded from cisbp.ccbr.utoronto.ca/bulk.php). The site position was identified as the mean of 'Start' and 'End'. TFs with less than 10,000 sites on autosomes were excluded. For each remaining TF, the top 10,000 sites were selected by choosing those with the highest 'peak.count' (number of times that peak has been observed across all experiments).

*DNase I hypersensitivity site selection*

DNase I hypersensitivity sites for a variety of tissue types were downloaded from zenodo.org/record/3838751/files/DHS_Index_and_Vocabulary_hg38_WM20190703.txt.gz (Meuleman, W. et al. Index and biological spectrum of human DNase I hypersensitive sites. Nature 584, 244–251 (2020)). These sites were split by tissue type for a total of 16 site lists. The 'summit' column was used as the site position. The sites were sorted by the number of samples where that site had been observed ('numsamples') and the top 10,000 most frequently observed sites were selected for each tissue type.

*ATAC-seq site selection for ER subtyping*

Assay for transposase-accessible chromatin using sequencing (ATAC-seq) site accessibility for primary breast cancer samples from The Cancer Genome Atlas (TCGA) were downloaded from the TCGA ATAC-seq hub (gdc.cancer.gov/about-data/publications/ATACseq-AWG) (Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. Science 362, eaav1898 (2018)). A file containing raw counts for all cancer type specific peaks were downloaded ('All cancer type-specific count matrices in raw counts') and the file containing breast cancer specific peaks was used ('BRCA_raw_counts.txt'). The locations of these sites and patient metadata were obtained from the supplemental tables in the paper (Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. Science 362, eaav1898 (2018)). Sites on autosomes were kept for further analysis for a total of 211,938 sites. Differentially accessible sites between ER+ (n=44) and ER- (n=15) tumors were identified using the DESeq2 software (Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550 (2014)). The software was run using default settings described in the 'quick start' guide. A differential expression experiment was run using the 'DESeq' and 'results' functions followed by log fold change shrinkage using the 'lfcShrink' function. Sites with a q-value $<5*10^{-4}$ were selected. Additionally, selected sites were further filtered based on the log2 fold change between ER+ and ER- tumors. Sites with a log2 fold change >0.5 were classified as ER+ specific, while sites with a log2 fold change <-0.5 were classified as ER- specific. These site lists were further split into sites shared with hematopoietic cells and those not shared with hematopoietic cells. Hematopoietic sites were obtained from a database of single cell ATAC-seq data (Satpathy, A. T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nature Biotechnology 37, 925–936 (2019)) (GEO accession number: GSE129785, peak file available here: ftp.ncbi.nlm.nih.gov/geo/series/GSE129nnn/GSE129785/suppl/GSE129785%5FscATAC%2DHematopoiesis%2DAll%2Epeaks%2Etxt%2Egz). Hematopoietic peaks were lifted over to hg38 using the UCSC liftover command line tool and sites that changed size during liftover (0.2% of peaks) were discarded. BRCA ATAC-seq sites that overlapped with Hematopoietic sites (Overlapping peaks were defined as site centers being within 500bp of one another) this was performed using pybedtools intersect (Dale, R. K., et al. Bioinformatics 27, 3423–3424 (2011); Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010)). This resulted in a total of 4 differential site lists: ER positive sites that were not shared with hematopoietic cells (18,240 sites), ER positive sites that were shared with hematopoietic cells (9,930 sites), ER negative sites that were not shared with hematopoietic cells (19,347 sites), and ER negative sites that were shared with hematopoietic cells (22,365 sites).

We then overlapped these differential ATAC-seq site lists with the top 10,000 sites for each of 338 transcription factors (TFs) using pybedtools intersect. An overlapping pair of sites was defined as having <500bp between site centers. Each differential ATAC-seq site list was compared against each list of TFBSs and the total number of ATAC sites overlapping one or more TFBS on the given list was recorded.

*Assessment of Griffin before and after GC correction*

Tumor fraction correlations at TFBS

For 191 MBC ULP samples with >0.1 tumor fraction, nucleosome profiling with and without GC correction was performed on the top 10,000 sites for each of 377 transcription factors (TFs). For each TF, the relationship between central coverage and tumor fraction was modeled using scipy.stats.linregress (Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17, 261–272 (2020)) producing a Pearson correlation (r) and line of best fit. Root mean squared error (RMSE) was calculated from the line of best fit. This was performed both before and after GC correction as illustrated for Lyl-1 in Fig. 2e. For all 377 TFs, the RMSE values before and after GC correction were compared using a Wilcoxon signed-rank test (two-sided).

Mean absolute deviation (MAD) at TFBS

For 215 healthy donors, nucleosome profiling with and without GC correction was performed on the top 10,000 sites for each of 377 TFs. For each TF, the MAD of the central coverage values was calculated both before and after GC correction. For all 377 TFs, the MAD values before and after GC correction were compared using a Wilcoxon signed-rank test (two-sided).

*Machine learning, bootstrapping, and performance evaluation procedure*

To detect cancer, predict tissue type, or predict ER subtype, we used logistic regression with Ridge regularization (i.e. L2 norm) as implemented in scikit-learn (Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)). All feature values were scaled to a mean of 0 and a standard deviation of 1 prior to performing bootstrapping and fitting the models. We used the following bootstrapping procedure to train and assess the performance of our models. First, we selected n samples with replacement from the full set of n samples and used this as a training set. Samples that were not selected were used as the test set. We then used 10-fold cross-validation on the training set to select the parameter 'C' (inverse of the regularization strength) from the following options: $10^{-5}$, $10^{-4}$, $10^{-3}$, $10^{-2}$, $10^{-1}$,$10^{0}$, $10^{1}$, $10^{2}$. To account for class imbalances in the data we used set the 'class weight' parameter to 'balanced' to adjust the sample weighs inversely proportional to the class frequencies. We trained a final model on all the training data using the selected regularization strength. Finally, we tested this model on the test set and recorded the performance (accuracy and AUC values) and probabilities from each sample. Then, a new training set was selected, and the procedure was repeated for 1000 iterations (for cancer detection and tissue of origin analysis) or 1000 iterations (for breast cancer subtyping). After completing the bootstrap iterations, we calculated the AUC and accuracy from each bootstrap iteration and used these to generate the mean and 95% confidence interval around each of these values. To visualize the mean ROC curve, we used the median probability from all bootstraps where that sample was included in the test set. For further downstream analyses, including the comut plot barplots and timelines we used this same median probability.

*Features used for cancer detection classification*

To detect cancer, we applied the logistic regression approach described above and built four different models using four different sets of features extracted from the pan cancer patient samples and healthy donor samples. First, we performed nucleosome profiling in these samples (selecting fragments 100-200bp in length) on the 377 selected TFs from the GTRD database. We extracted three features (as described above) from each coverage profile for a total of 1,014 features. We reduced the dimensionality of the data using PCA and selected the features that explained 80% of the variance. These PCA components were then used as the inputs for the logistic regression model.

Second, we performed nucleosome profiling on these same samples and sites but selected only 'short' fragments (35-150bp) to be counted in the nucleosome profiles.

Third, we downsampled these samples to ~0.1x coverage (procedure described below) and performed nucleosome profiling for the same 377 TFs selecting fragments 100-200bp in length.

Fourth, we used the original (not downsampled) samples and performed nucleosome profiling at the 16 tissue-specific DHS site lists described above. We extracted the same 3 features from each site profile for a total of 48 features.

*Downsampling of pan-cancer and healthy donor cfDNA sequencing data*

1-2x WGS of pan-cancer patient and healthy donor bam files aligned to hg38 were downsampled using Picard DownSampleSam. The probability used by DownSampleSam was calculated based on a target of 2,463,109 read pairs which resulted in approximately 0.11x coverage as calculated by Picard CollectWgsMetrics.

Downsampled bam files were realigned to hg19 for use in the Ulz pipeline. The realignment procedure was the same as above but using the hg19 genome (downloaded from hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.gz) and hg19 known polymorphic sites for base recalibration (downloaded from gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg37/Mills_and_1000G_gold_standard.indels.hg37.vcf.gz and ftp.ncbi.nih.gov/snp/organisms/human_9606_b151_GRCh37p13/VCF/GATK/All_20180423.vcf.gz).

*ER status classification in the MBC cohort*

To predict ER status, we applied the logistic regression approach described above to features extracted from the MBC patient samples. Because some patients had multiple samples, we modified the bootstrapping procedure to select 139 patients (rather than samples) with replacement from a full set of 139 patients. For each selected patient, all samples from that patient were added to the training set (If a patient was selected multiple times, all their samples were included multiple times). This ensured that separate samples from the same patient (biological replicates) could not appear in both the training and test set. Samples from patients that were not selected were used as the test set.

Using these training and tests sets, we built three different models based on three different sets of features. First, we applied nucleosome profiling using 100-200bp fragments to the 377 TFs from GTRD and extracted 3 features per profile for a total of 1131 features. We then used PCA to identify the components that explained 80% of the variance as described above. Second, we applied nucleosome profiling using 100-200bp fragments to the 4 ER differential ATAC seq lists and extracted 3 features per profile for a total of 12 features. Lastly, we applied nucleosome profiling using 35-150bp fragments to the 4 ER differential ATAC seq lists and extracted 3 features per list for a total of 12 features.

For evaluating the models, we only included the first timepoint for each patient in the test set when calculating the accuracy and AUC for each bootstrap iteration. This prevented a small number of patients with many samples from having a large impact on the scores.

*Transcription factor profiling using pipeline from Ulz et al.*

We downloaded the Transcription Factor Profiling pipeline published by Ulz and colleagues from Github (github.com/PeterUlz/TranscriptionFactorProfiling) (Ulz, P. et al.

Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nature Communications 10, 4666 (2019)) and ran it using the following procedure as described in the paper. hg19 aligned bam files were used because the pipeline was written to for this version of the genome. Scripts were modified so that they worked in python3. We trimmed the reads in each bam to 60bp using 'trim from bam single end' with modifications to skip unaligned reads. We ran ichorCNA on the original (untrimmed) bam using the default ichorCNA settings for hg19 except the bin size, which was modified to 50,000bp and no panel of normals. We then ran the transcription factor profiling analysis on the trimmed bam using the script run_tf_analyses_from_bam.py with options '-calccov' and '-a tf_gtrd_1000sites' and the ichorCNA corrected depth file as the '-norm-file'. This ran transcription factor profiling on 1,000 sites for each of 504 TFs. Finally, we ran the scoring pipeline. We used the high frequency amplitude ('HighFreqRange') for each of the 504 TFs in the accessibility output file (Accessibility1KSitesAdjusted.txt) as the features for a logistic regression model using the same bootstrapping scheme described above.

*Data availability*

Sequencing data used in this study was obtained from dbGaP (accession phs001417.v1.p1) and EGA (dataset ID EGAD00001005339).

*Code availability*

Griffin software and the subtype classifier tool can be obtained from github.com/adoebley/Griffin. Code for analysis and machine learning models can be accessed at github.com/adoebley/Griffin_analyses.

## Example 2

Example 1 above is a proof-of-concept demonstration that sequence analysis applying an embodiment of the Griffin workflow can enhance sequence signals with sufficient power and specificity to allow determination of breast cancer subtypes from low pass sequencing data. This Example expands the application of Griffin workflow to other cancer types and makes use of data from an alternative sequence profiling platform. Specifically, histone modification profiling was performed using the CUT & RUN on different subtypes of prostate cancer cells. As with Example 1, the Griffin workflow provided robust signals to clearly differentiate different subtypes of prostate cancer, demonstrating the power and flexibility of the analytic workflow.

Background

Metastatic castration-resistant prostate cancer (mCRPC) describes the stage in which the disease has developed resistance to androgen ablation therapies and is lethal. Androgen receptor signaling inhibitors (ARSI), designed for the treatment of CRPC, repress androgen receptor (AR) activity and improve survival, but these therapies eventually fail. Since the adoption of ARSI as standard-of-care for mCRPC, there has been a prominent increase in the frequency of treatment-resistant tumors with neuroendocrine (NE) differentiation and features of small cell carcinomas. These aggressive tumors may develop through a resistance mechanism of trans-differentiation from AR-positive adenocarcinoma (ARPC) to NE prostate cancer (NEPC) that lack AR activity. Additional phenotypes can also arise based on expression of AR activity and NE genes, including AR-low prostate cancer (ARLPC) and double-negative prostate cancer (DNPC; AR-null/NE-null). Distinguishing prostate cancer subtypes has clinical relevance in view of differential responses to therapeutics, but the need for a biopsy to diagnose tumor histology can be challenging: invasive procedures are expensive and accompanied by morbidity, a subset of tumors are not accessible to biopsy, and bone sites pose particular challenges with respect to sample quality.

Circulating tumor DNA (ctDNA) released from tumor cells into the blood as cell-free DNA (cfDNA) is a non-invasive "liquid biopsy" solution for accessing tumor molecular information. The analysis of ctDNA to detect mutation and copy-number alterations has served to classify genomic subtypes of CRPC tumors. However, the defining losses of *TP53* and *RB1* in NEPC do not always lead to NE trans-differentiation. Rather, ARPC and NEPC tumors are associated with distinct reprogramming of transcriptional regulation. Methylation analysis of cfDNA in mCRPC to profile the epigenome shows promise for distinguishing phenotypes, but requires specialized assays such as bisulfite treatment, enzymatic treatment, or immunoprecipitation.

The majority of cfDNA represents DNA protected by nucleosomes when released from dying cells into circulation, leading to DNA fragmentation that is reflective of the non-random enzymatic cleavage by nucleases. Emerging approaches to analyze cfDNA fragmentation patterns from plasma for studying cancer can be performed directly from standard whole genome sequencing (WGS). cfDNA fragments have the characteristic size of 167 bp, consistent with protection by a single core nucleosome octamer and histone linkers, but the size distribution may vary between healthy individuals and cancer patients. Recent studies have demonstrated that the nucleosome occupancy in cfDNA at the transcription start site (TSS) and transcription factor binding site (TFBS) can be used to infer gene expression and transcription factor (TF) activity from cfDNA. However, nucleosome positioning and spacing are dynamic in active and repressed gene regulation. A detailed understanding of the nucleosome organization and positioning patterns associated with transcriptional regulation has not been fully explored in cfDNA.

A major challenge for ctDNA analysis is the low tumor content (tumor fraction) in patient plasma samples. By contrast, plasma from patient-derived xenograft (PDX) models may contain nearly pure human ctDNA after bioinformatic exclusion of mouse DNA reads. This provides a resource that is ideal for studying the properties of ctDNA, developing new analytical tools, and validating both genetic and phenotypic features by comparison to matching tumors. In this study, WGS of ctDNA from mouse plasma across 24 CRPC PDX lines with diverse phenotypes was performed deep. Applying the computational methods described in Example 1, the nucleosome patterns in was comprehensively interrogated across genes, regulatory loci, TFBSs, TSSs, and open chromatin sites to reveal transcriptional regulation associated with mCRPC phenotypes. Finally, a probabilistic model was designed to accurately classify treatment-resistant tumors into divergent phenotypes and validated its performance in 159 plasma samples from three mCRPC patient cohorts. Overall, these results highlight that transcriptional regulation of tumor phenotypes can be ascertained from ctDNA and has potential utility for diagnostic applications in cancer precision medicine.

Results

*Comprehensive resource of matched tumor and liquid biopsies from patient derived xenograft (PDX) models of advanced prostate cancer*

Twenty-six models from the LuCaP PDX series of advanced prostate cancer with well-defined mCRPC phenotypes were used (Nguyen et al., (2017). LuCaP Prostate Cancer Patient-Derived Xenografts Reflect the Molecular Heterogeneity of Advanced Disease and Serve as Models for Evaluating Cancer Therapeutics. The Prostate *77*, 654–671). The models consisted of 18 classified as ARPC, two classified as AR-low and NE-negative prostate cancer (ARLPC), and six classified as NEPC (FIG. 11A). For each PDX line, mouse plasma was pooled from seven to ten mice, cfDNA was extracted, and deep whole genome sequencing was performed (WGS; mean 38.4x coverage, range 21 – 85x) (Methods, FIG. 11A). Twenty-five lines had human ctDNA comprising more than 10%

of the sample (mean 52.9%, range 10.6 – 96%) with NEPC samples having significantly higher human fractions (mean 85.1%, range 77.1 – 96%, two-tailed Mann-Whitney U test $p = 9.6 \times 10^{-4}$) (FIG. 11B). Bioinformatic subtraction of mouse sequenced reads was used to obtain nearly pure human ctDNA data (Methods). After subsequent filtering by human ctDNA sequencing coverage, 24 PDX lines remained for further analysis (16 ARPC, 6 NEPC, 2 ARLPC; mean 20.5x, range 3.8 – 50.6x). In the matching tumors, Cleavage Under Targets and Release using Nuclease (CUT&RUN) was performed to profile H3K27ac, H3K4me1, and H3K27me3 histone post-translational modifications (PTMs) (Meers et al., (2019). Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. Epigenetics & Chromatin *12*, 42; Skene and Henikoff (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. ELife *6*, e21856.) It was hypothesized that nucleosome organization inferred from ctDNA reflects the transcriptional activity state regulated by histone PTMs (Zhou et al. (2011). Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet *12*, 7–18).

To study transcriptional regulation in mCRPC phenotypes from ctDNA, four different features were interrogated: local promoter coverage, nucleosome positioning, fragment size analysis, and composite TFBSs and open chromatin sites analysis using the Griffin framework (Example 1; and Doebley et al. (2021). Griffin: Framework for clinical cancer subtyping from nucleosome profiling of cell-free DNA. MedRxiv 2021.08.31.21262867) (FIG. 11A, Methods). Three different local regions within ctDNA were analyzed: all gene promoters and gene bodies and sites of histone PTMs guided by CUT&RUN analysis. Next, ctDNA was analyzed at transcription factor binding sites (TFBSs) and open chromatin regions. For each transcription factor (TF), ctDNA coverage at TFBSs were aggregated into composite profiles representing the inferred activity (Example 1 and Doebley et al., 2021; Ulz et al. (2019). Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nature Communications *10*, 4666). Similarly, features in the composite profiles of subtype-specific open chromatin regions were extracted for analyzing the signatures of chromatin accessibility in ctDNA. Altogether, a multi-omic sequencing dataset was assembled from matching tumor and plasma for a total of 24 PDX lines, making this a unique molecular resource and platform for developing transcriptional regulation signatures of tumor phenotype prediction from ctDNA.

*Characterizing transcriptional activity of AR and ASCL1 in PDX phenotypes through analysis of tumor histone modifications and ctDNA*

Prostate cancer phenotypes in mCRPC patients have distinct transcriptional signatures and these are also observed in the LuCaP PDX lines (Labrecque et al. (2021b). RNA Splicing Factors SRRM3 and SRRM4 Distinguish Molecular Phenotypes of Castration-Resistant Neuroendocrine Prostate Cancer. Cancer Research *81*, 4736–4750). The transcriptional activity was further characterized in different tumor phenotypes by studying epigenetic regulation via histone PTMs. Broad peak regions for H3K4me1 (median of 17,643 regions, range 1,894 – 64,934), H3K27ac (median 7,093, range 1610 - 34,047), and H3K27me3 (median 8,737, range 2,024 - 42,495) were identified in the tumors of the 24 PDX lines and an additional nine LuCaP PDX lines where only tumor was available (total of 25 ARPC, 2 ARLPC, and 6 NEPC) (Methods). Using unsupervised clustering and principal components analysis (PCA), putative active regulatory regions of enhancers and promoters (H3K27ac, H3K4me1) and gene repressive heterochromatic mark (H3K27me3) were identified that were specific to ARPC, ARLPC, and NEPC phenotypes (Soares et al. (2017). Determinants of Histone H3K4 Methylation Patterns. Molecular Cell *68*, 773-785.e6).

AR and ASCL1 are two key differentially expressed TFs with known regulatory roles in ARPC and NEPC phenotypes, respectively (Brady et al. (2021). Temporal evolution of cellular heterogeneity during the progression to advanced AR-negative prostate cancer. Nat Commun *12*, 3372; Cejas et al. (2021). Subtype heterogeneity and epigenetic convergence in neuroendocrine prostate cancer. Nat Commun *12*, 5775; Rapa et al. (2008). Human ASH1 expression in prostate cancer with neuroendocrine differentiation. Mod Pathol *21*, 700–707; Wang et al. (2020). Molecular tracing of prostate cancer lethality. Oncogene *39*, 7225–7238). When inspecting AR binding sites in ARPC tumors, increased signals were observed from flanking nucleosomes with H3K27ac PTMs compared to the other phenotypes (area under mean peak profile of 18.46 vs. 15.08 in ARLPC and 10.63 in NEPC) (FIG. 12A, Methods). The strongest signals were also observed at the nucleosome depletion region (NDR) in ARPC for H3K27ac (1.54 coverage decrease vs. 0.78 for ARLPC and 0.41 for NEPC). Conversely, in NEPC tumors, stronger signals were observed at nucleosomes with H3K27ac PTMs flanking ASCL1 binding sites (area under mean peak profile 62.65 vs. 29.18 for ARLPC and 10.83 for ARPC), and stronger NDR signals (2.26 coverage decrease vs. 0.19 for

ARPC and 0.37 for ARLPC). Similar trends were observed for H3K4me1 PTMs in the LuCaP PDX lines.

The ctDNA composite coverage profiles were analyzed at TFBSs to evaluate the nucleosome accessibility, whereby lower normalized central (±30 bp window) mean coverage across these sites suggests more nucleosome depletion (Methods). For AR TFBSs, the strongest signal for nucleosome depletion was observed in ARPC, as indicated by the lowest mean central coverage (average 0.64, n=16), compared to moderate signals for ARLPC (average 0.88, n=2), and weakest signals for NEPC (average 0.95, n=6) (FIG. 12B). Conversely, the composite coverage profile at ASCL1 TFBSs showed the strongest nucleosome depletion for NEPC samples (mean central coverage 0.69) compared to ARLPC (0.86) and ARPC (0.88) (FIG. 12C). These observations were consistent with the differential binding activity by AR and ASCL1 in their respective phenotypes from tumor tissue. Furthermore, the ctDNA coverage patterns of the nucleosome depletion in ctDNA resembled the NDR flanked by nucleosomes with H3K27ac and H3K4me1 peak profiles, which was exemplified when analyzing only nucleosome-sized fragments (140 bp – 200 bp) generated by CUT&RUN (FIG. 12A). Together, these results suggest that the nucleosome depletion in ctDNA at AR and ASCL1 binding sites represents active TF binding and regulatory activity in specific prostate PDX tumor phenotypes.

*Nucleosome patterns at gene promoters inferred from ctDNA are consistent with transcriptional activity for phenotype-specific genes*

Forty-seven genes comprising 12 ARPC and 35 NEPC lineage markers established previously (Beltran et al. (2016). Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. Nature Medicine *22*, 298–305; Labrecque et al. (2021b). RNA Splicing Factors SRRM3 and SRRM4 Distinguish Molecular Phenotypes of Castration-Resistant Neuroendocrine Prostate Cancer. Cancer Research *81*, 4736–4750) were selected and confirmed by differential expression analysis from PDX tumor RNA-Seq data (FIG. 12D, Methods). To assess the activity of these genes from ctDNA, the ctDNA fragment size in TSSs (± 1 kb window) and gene bodies were analyzed. It was found that the differential size variability between phenotypes was positively correlated with relative expression (Spearman's $\rho = 0.844$, $p = 9.4 \times 10^{-14}$, FIG. 12E, Methods). Next, the relative ctDNA coverage at the TSS (± 1 kb) was analyzed but no association between the phenotypes was observed. However, closer inspection of the ctDNA coverage patterns at the promoters revealed consistent nucleosome organization for transcription activity and repression (Jiang and Zhang (2021). On the role of transcription in positioning nucleosomes. PLOS Computational Biology *17*, e1008556; Klemm et al. (2019). Chromatin accessibility and the regulatory epigenome. Nature Reviews Genetics *20*, 207–220; Oruba et al. (2020). Role of cell-type specific nucleosome positioning in inducible activation of mammalian promoters. Nat Commun *11*, 1075; Ramachandran et al. (2017). Transcription and Remodeling Produce Asymmetrically Unwrapped Nucleosomal Intermediates. Molecular Cell *68*, 1038-1053.e4) (FIG. 12D). Therefore, the genes were grouped based on differential signals in H3K27me3 histone PTMs, which are associated with repressed transcription or heterochromatic compaction. For 25 genes (Group 1) without differential H3K27me3 peaks, including AR, FOXA1, KLK3 and ASCL1, nucleosome depletion was observed at the TSS consistent with presence of active PTMs, such as for AR (mean coverage 0.47, n=16) in ARPC and ASCL1 (0.30, n=6) in NEPC samples (FIG. 12F). By contrast, increased coverage was observed at the TSS of AR (1.08) in NEPC and ASCL1 (0.42) in ARPC, which supports the nucleosome depletion in the absence of PTMs and inactive transcription. For 22 genes (Group 2) with differential H3K27me3 peaks, including STEAP1, CHGB and SRRM4, a relatively more consistent increase was observed in nucleosome occupancy and phasing in the TSS as well as in the gene body for NE-specific genes (FIG. 12G). The neural signaling genes in this group, such as UNC13A and INSM1, had reduced signals for nucleosome positioning, consistent with the heterogeneous ('fuzzy') nucleosome patterns described for actively transcribed genes (Jiang and Pugh (2009). Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet *10*, 161–172; Lai and Pugh (2017). Understanding nucleosome dynamics and their links to gene expression and DNA replication. Nat Rev Mol Cell Biol *18*, 548–562). Interestingly, while UNC13A was repressed in ARPC tumors, it did not have H3K27ac nor H3K4me1 accessible PTM marks in NEPC tumors despite being expressed. These results illustrate that ctDNA analysis can reveal patterns that are consistent with transcriptional regulation by histone modifications for key genes that define prostate cancer phenotypes.

*Inferred TF activity from analysis of nucleosome accessibility at TFBSs in ctDNA confirms key regulators of tumor phenotypes*

To characterize the lineage-defining TFs in prostate tumor phenotypes, nucleosome accessibility was considered at TFBSs in PDX ctDNA. 107 TFs were identified based on the intersection of 338 TFs analyzed using Griffin and 404 differentially expressed TFs between ARPC and NEPC PDX tumors (Methods). Of these TFs, 38 had significantly different accessibility in ctDNA between ARPC and NEPC phenotypes (two tailed Mann-Whitney U test, Benjamini-Hochberg adjusted $p < 0.05$). Through unsupervised hierarchical clustering of composite TFBS central coverage values for the 107 TFs, distinct groups of TFs were observed in PDX ctDNA (FIG. 13). REST had the largest difference in accessibility as supported by a decrease in coverage within ARPC models compared to NEPC ($\log_2$ fold-change -0.77, adjusted $p = 5.7 \times 10^{-4}$). FOXA1, and GRHL2 were significantly more accessible in ARPC (and ARLPC) samples compared to NEPC ($\log_2$ fold-change < -0.57, adjusted $p < 1.3 \times 10^{-3}$). AR, HOXB13, and NKX3-1 had higher accessibility in ARPC compared to NEPC ($\log_2$ fold-change < -0.37, adjusted $p < 1.3 \times 10^{-3}$), but with only moderate accessibility in ARLPC, as expected. Interestingly, progesterone receptor (PGR) also had high accessibility in ARPC ($\log_2$ fold-change -0.33, adjusted $p = 2.6 \times 10^{-3}$). A group of ARPC-regulated genes that followed a similar trend were also observed, including the glucocorticoid receptor (NR3C1) and other nuclear hormone receptors (NR2F2, RARG), pioneer factors GATA2 and GATA3, and nuclear factors HNF4G and HNF1A ($\log_2$ fold-change < -0.10, adjusted $p < 0.027$).

For factors that had higher accessibility in NEPC models compared to ARPC and ARLPC, ASCL1 had the largest TFBS coverage difference ($\log_2$ fold-change 0.36, adjusted $p = 5.7 \times 10^{-4}$, FIG. 12C, FIG. 13F). Other TFs, including RUNX1, BCL11B, POU3F2, NEUROG2, and SOX2 also had higher activity in NEPC ($\log_2$ fold-change > 0.06, adjusted $p < 0.048$), although the difference was modest. HEY1, IRF1, and IKZF1 had a similar trend consistent with increased accessibility in NEPC samples but were not significantly different from ARPC (adjusted $p > 0.10$). While NKX2-1 and CEBPA had increased accessibility in NEPC compared to ARPC (although not significant with adjusted p = 0.47 and 0.36 respectively), these factors were also modestly active in ARLPC. Other notable factors such as MYC and ETS transcription family genes (ETV4, ETV5, ETS1, ETV1) had high accessibility across all phenotypes, while NEUROD1, RUNX3, and TP63 were inaccessible in nearly all samples. Overall, the accessibility of known prostate cancer regulators were identified, including ASCL1, NR3C1, HNF4G, HNF1A, and SOX2 (Arora et al. (2013). Glucocorticoid Receptor Confers Resistance to Antiandrogens by Bypassing Androgen Receptor Blockade. Cell *155*, 1309–1322; Mu et al. (2017). SOX2 promotes lineage plasticity and antiandrogen resistance in TP53- and RB1-deficient prostate cancer. Science *355*, 84–88; Shukla et al. (2017). Aberrant Activation of a Gastrointestinal Transcriptional Circuit in Prostate Cancer Mediates Castration Resistance. Cancer Cell *32*, 792-806.e7), that have not been shown before from ctDNA analysis in these tumor phenotypes.

*Phenotype-specific open chromatin regions in PDX tumor tissue are reflected in ctDNA profiles of nucleosome accessibility*

Nucleosome profiling from cfDNA sequencing analysis has shown agreement with overall chromatin accessibility in tumor tissue (Snyder et al. (2016). Cell-free DNA Comprises an in Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell *164*, 57–68; Sun et al. (2019). Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. Genome Research *29*, 418–427; Ulz et al. (2019). Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nature Communications *10*, 4666); however, its application for distinguishing tumor phenotypes has been limited. the use of ATAC-Seq data from tumor tissue for 10 LuCaP PDX lines (5 ARPC and 5 NEPC) was investigated to inform phenotype differences in chromatin accessibility (Cejas et al. (2021). Subtype heterogeneity and epigenetic convergence in neuroendocrine prostate cancer. Nat Commun *12*, 5775). An initial set of 28,765 ARPC and 21,963 NEPC differential consensus open chromatin regions was defined, which was further restricted to those that overlapped TFBSs for 338 TFs, resulting in 15,881 ARPC and 11,694 NEPC sites (Methods, FIG. 14A). For ARPC-specific open chromatin sites, decreased overall composite site coverage (+/- 1 kb window) and central coverage (+/- 30 bp) were observed in the ctDNA for ARPC PDX lines (mean central coverage 0.75, n=16) compared to NEPC lines (mean 0.96, n=6) and cfDNA from healthy human donors (mean 0.97, n=14) (FIG. 14B). Conversely, for NEPC-specific open chromatin sites, coverage was decreased in ctDNA for NEPC lines (mean 0.89) compared to ARPC lines (mean 1.01) and healthy donors cfDNA (mean 1.00) (FIG. 14C). These results confirmed that tumor tissue chromatin accessibility can be corroborated in ctDNA and that ARPC and NEPC phenotypes have distinct ctDNA composite site coverage profiles.

*Comprehensive evaluation of ctDNA features across genomic contexts for CRPC phenotype classification*

To assess the utility of ctDNA nucleosome profiling for informing prostate cancer phenotype classification, groups of global genome-wide ctDNA features were systematically evaluated: fragment sizes, local coverage profiling, and composite site coverage profiling (FIG. 11A). From principal components analysis (PCA), distinct feature signals were observed between ARPC and NEPC phenotypes for composite TFBS coverage of TFs and fragment size variability at global sites of PTMs (FIG. 14D, Methods). In addition to these features, similar approaches previously reported were also included, including short-long fragment ratio and local coverage patterns at the TSS (max wave height between -120bp to 195bp) (Cristiano et al. (2019). Genome-wide cell-free DNA fragmentation in patients with cancer. Nature *570*, 385–389; Ulz et al. (2016b). Inferring expressed genes by whole-genome sequencing of plasma DNA. Nature Genetics *48*, 1273–1278) (Methods).

All combinations of coverage and fragment size features were quantitatively evaluated for different genomic contexts to investigate their potential to classify ARPC and NEPC phenotypes. For each feature set, 100 iterations of stratified cross-validation were conducted using a supervised machine learning classifier (XGBoost) on ctDNA samples from the 16 ARPC and 6 NEPC models and computed the area under the receiver operating characteristic curve (AUC) (Methods). First, an established set of 10 genes associated with AR activity was evaluated (Bluemn et al. (2017). Androgen Receptor Pathway-Independent Prostate Cancer Is Sustained through FGF Signaling. Cancer Cell *32*, 474-489.e6; Labrecque (2021a). The heterogeneity of prostate cancers lacking AR activity will require diverse treatment approaches. Endocrine-Related Cancer *28*, T51–T66). It was observed that the phased nucleosome distance at H3K27ac sites and the central coverage at TSSs had moderate predictive performance (AUC 0.88). When considering all PTM sites, promoters, genes, TFs, and open chromatin regions, the best performing features included mean fragment size at H3K4me1 sites (n=9,750, AUC 1.0) and promoter TSSs (n=17,946, AUC 1.0), and both open chromatin composite site features (AUC 1.0) (FIG. 14E).

*Accurate classification of ARPC and NEPC phenotypes from patient plasma using a probabilistic model informed by PDX ctDNA analysis*

An important consideration and challenge in analyzing plasma from patients is the presence of cfDNA released by hematopoietic cells, which leads to a lower ctDNA fraction (i.e., tumor fraction). Furthermore, the small patient cohorts with available tumor phenotype information make supervised machine learning approaches suboptimal. Therefore, a probabilistic model was developed to estimate the proportion of ARPC and NEPC from an individual plasma sample, accounting for the tumor fraction (**Methods**). A focused was made on the phenotype-specific open chromatin composite site features and the PDX plasma ctDNA signals were used (FIGS. 14B and 14C) to inform the model. The model produces a normalized prediction score that represents the estimated signature of ARPC (lower values) and NEPC (higher values). This method was applied to benchmarking datasets generated by simulating varying tumor fractions and sequencing coverages using five ARPC and NEPC PDX ctDNA samples each (FIG. 14F, Methods). A 1.0 AUC was achieved at 25X coverage down to 0.01 tumor fraction, 1.0 AUC at 1X down to 0.2 tumor fraction, and 1.0 AUC at 0.2x coverage at 0.3 tumor fraction, suggesting a possible upper-bound performance for classifying samples with lower tumor fraction in plasma (FIG. 14G).

To test the classification performance of the model on patient samples, a published dataset of ultra-low-pass whole genome sequencing (ULP-WGS) of plasma cfDNA (mean coverage 0.52X, range 0.28-0.92X) from 101 mCRPC patients comprising 80 adenocarcinoma (ARPC) and 21 NEPC samples (DFCI cohort I) was analyzed (Berchuck et al. (2022). Detecting Neuroendocrine Prostate Cancer Through Tissue-Informed Cell-Free DNA Methylation Analysis. Clinical Cancer Research *28*, 928–938). Using the model, which was unsupervised and used parameters informed only by the PDX analysis, an overall AUC of 0.96 was achieved (FIG. 15A). When considering samples with high ($\geq 0.1$) and low ($< 0.1$) tumor fraction, the model had an 0.97 AUC and 0.76 AUC, respectively. An optimal overall performance at 97.5% specificity (ARPC) and 90.4% sensitivity (NEPC) was identified, which corresponded to the prediction score of 0.3314 (FIG. 15A). In another published dataset of 11 mCRPC samples from 6 patients who had high PSA, treatment with ARSI, or both (DFCI cohort II) (Choudhury et al. (2018). Tumor fraction in cell-free DNA as a biomarker in prostate cancer. JCI Insight *3*; Viswanathan et al., 2018), the model correctly classified patients as ARPC in 11 (100%) WGS (~20x) and 8 (73%) ULP-WGS (~0.1x) samples when using the optimal score cutoff (FIG. 15B).

Next, 61 clinical plasma samples from 30 CRPC patients with ARPC, NEPC, and mixed phenotypes that are representative of typical clinical histories were analyzed. ULP-WGS of cfDNA was performed and 47 samples from 30 patients (26 ARPC, 5 NEPC, and 16 mixed phenotypes) were selected based on tumor fraction and AR copy number status for deeper WGS (mean 22.13X coverage, range 15.15X – 31.79X) (Methods). For the 26 samples with ARPC clinical phenotype, all were predicted to be predominantly ARPC using the score cutoff of 0.3314 (FIG. 15C). For NEPC clinical phenotype, all five were predicted to be NEPC with scores above the cutoff. A negative association was also noted between the patient ctDNA coverage at open chromatin sites and the tumor fraction for both ARPC (Spearman's ρ = -0.93) and NEPC predictions (Spearman's ρ = -1.00), suggesting that the observed ctDNA signals were likely tumor-specific. From ULP-WGS data, 22 (84%) samples with ARPC clinical phenotype and all five (100%) samples with NEPC clinical phenotype were correctly predicted (FIG. 15C). The remaining 16 samples had clinical histories or tumor histologies that reflected mixed phenotypes such as a tumor with AR-positive adenocarcinoma intermixed with NEPC (FIG. 15C). For 12 samples that included presence of ARPC in the mixed clinical phenotype, 10 (83%) were classified as ARPC at the optimal score cutoff. For all three samples that had presence of NEPC but no ARPC in the clinical phenotype, the model classified them as NEPC. Overall, an accuracy of 100% for was achieved WGS (87% for ULP-WGS) data for samples with unambiguous clinical phenotypes. However, the variable predictions for mixed or ambiguous phenotypes underscore the complexities associated with classification in patients with advanced prostate cancer where tumor heterogeneity can be observed.

Discussion

The study presented here is believed to be the largest sequencing study to date of human ctDNA from mouse plasma of PDX models. The sequencing of mouse plasma provided a unique opportunity to comprehensively interrogate the epigenetic nucleosome patterns in ctDNA from well-characterized tumor models. Computational methodologies were developed and applied to construct a multitude of ctDNA features, each of which were associated with the transcriptional regulation in the LuCaP PDX models across CRPC tumor phenotypes. Using features learned from the PDX ctDNA, a probabilistic model was developed to accurately classify ARPC and NEPC phenotypes from patient plasma in three clinical cohorts.

The use of PDX mouse plasma overcomes the challenge of low ctDNA content or incomplete knowledge of the tumor when studying patient samples and can expedite development of cfDNA diagnostics, basic cancer research, and clinical translation.

Furthermore, the LuCaP ctDNA sequencing data complements the maturing characterization of CRPC tumor phenotypes from tissue. In addition to supporting molecular studies of CRPC, the ctDNA data and the disclosed approaches expand on the potential utility of PDX models for translational research. While these data were focused on ARPC and NEPC phenotypes, this study can serve as a framework for the use of PDX plasma from additional CRPC phenotypes and other cancers models.

The analysis of the LuCaP PDX ctDNA sequencing data confirmed the activity of key regulators between ARPC and NEPC phenotypes, including a set of 47 established differentially expressed gene markers. While gene expression inference from ctDNA has been shown in proof-of-concept studies (Ulz et al. (2016b). Inferring expressed genes by whole-genome sequencing of plasma DNA. Nature Genetics *48*, 1273–1278; Zhu et al. (2021). Tissue-specific cell-free DNA degradation quantifies circulating tumor DNA burden. Nature Communications *12*, 2229), the PDX ctDNA allowed for a detailed dissection of nucleosome organization associated with transcriptional activity of individual genes that define the tumor phenotypes.

In addition to the existing molecular profiling available for these models, this study now provides characterization of histone PTMs in LuCaP PDX tumors using CUT&RUN. At regions with these PTMs on histone tails, nucleosome patterns inferred in ctDNA were observed that were consistent with active or repressed gene transcription. This is believed to be the first time that ctDNA analysis has been performed in the context of histone PTMs and provides a blueprint to develop new approaches for studying additional epigenetic alterations using PDX plasma.

While the regulation of key factors such as AR, HOXB13, NKX-3.1, FOXA1, and REST has been shown from ctDNA in CRPC (Ulz et al. (2019). Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nature Communications *10*, 4666), this study reveals the differential activity of other key factors in CRPC for the first time from ctDNA analysis. This included the glucocorticoid receptor (NR3C1), nuclear factors HNF4G and HNF1A, and pioneering factors GATA2 and GATA3, all of which are associated with prostate adenocarcinoma (ARPC) (Arora et al. (2013). Glucocorticoid Receptor Confers Resistance to Antiandrogens by Bypassing Androgen Receptor Blockade. Cell *155*, 1309–1322; Chaytor et al., 2019; Shukla et al., 2017). ASCL1 is a pioneer TF with roles in neuronal differentiation and was recently described to be active during NE trans-differentiation and in NEPC (Cejas et al., 2021;

Rapa et al., 2008). To our knowledge, this study is the first to demonstrate ASCL1 binding site accessibility and provide a detailed characterization of its transcriptional activity in NEPC from plasma ctDNA.

This study provides an expansive analysis of TFBSs for 338 factors in each plasma sample without the need for chromatin immunoprecipitation or other epigenetic assays. However, there was no observed significant difference in accessibility for 69 out of the 107 TFs in ctDNA, which may be consistent with TF activity not necessarily being correlated with its own expression levels (Corces et al. (2018). The chromatin accessibility landscape of primary human cancers. Science *362*). On the other hand, the accessibility of TFBSs may not necessarily indicate true TF activity, such as binding of multiple factors to the same locus. Moreover, this analysis was based on TFBSs obtained from public databases; however, prostate phenotype-specific TF cistromes may better guide this approach.

State-of-the-art computational approaches built on existing and new concepts of ctDNA data analysis were applied to extract tumor-specific features. Other approaches have also considered regions, such as TSSs, TFBSs, and DNase hypersensitivity sites (Peneder et al. (2021). Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden. Nat Commun *12*, 3230; Snyder (2016). Cell-free DNA Comprises an in Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell *164*, 57–68; Ulz et al. (2016b). Inferring expressed genes by whole-genome sequencing of plasma DNA. Nature Genetics *48*, 1273–1278; Ulz et al. (2019). Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nature Communications *10*, 4666); however, after a systematic evaluation, it was found that ctDNA features in open chromatin sites derived from ATAC-Seq of PDX tissue (Cejas et al. (2021). Subtype heterogeneity and epigenetic convergence in neuroendocrine prostate cancer. Nat Commun *12*, 5775) provided the highest performance for distinguishing CRPC phenotypes. An unsupervised probabilistic model is presented that estimates the proportion of ARPC and NEPC in patient plasma using a statistical framework informed by idealized parameters from the LuCaP PDX ctDNA analysis. This model does not require training on patient samples but does require tumor fraction estimates (ichorCNA (Adalsteinsson (2017). Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nature Communications *8*) and a prediction score cutoff determined from DFCI cohort I. The framework presented here can be extended to model multiple phenotype classes, provided the informative parameters for these additional states can be learned. Insights from additional datasets such as single-cell nucleosome and accessibility profiling (Fang et al. (2021). Comprehensive analysis of single cell ATAC-seq data with SnapATAC. Nat Commun *12*, 1337; Wu et al. (2021). Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. Nat Biotechnol *39*, 819–824) of PDX tumors and clinical samples may improve the resolution for ctDNA analysis.

Applying the prediction model to patient datasets with definitive clinical phenotypes yielded high performance despite using low depth of coverage sequencing. In particular, our performance for the DFCI cohort I was also consistent with the reported phenotype classification results using ctDNA methylation in the same patients (Berchuck et al. (2022). Detecting Neuroendocrine Prostate Cancer Through Tissue-Informed Cell-Free DNA Methylation Analysis. Clinical Cancer Research *28*, 928–938). Similarly, in the UW cohort, samples with well-defined clinical phenotypes had perfect concordance from deep WGS data. However, samples with mixed or ambiguous clinical phenotypes limited our ability to definitively assess the performance of the model because a subset of cases had complex clinical and histopathological features. Tumor heterogeneity and co-existence of different molecular phenotypes are common in mCRPC where treatment-induced phenotypic plasticity may vary within and between tumors in an individual patient. Larger studies with comprehensive assessment of the tumor histologies will be needed for developing future extensions of the model to predict mixed phenotypes from ctDNA.

In summary, this study illustrates for the first time that analysis of ctDNA from PDX mouse plasma at scale can facilitate a more detailed investigation of tumor regulation. These results, together with the suite of computational methods presented here, highlight the utility of ctDNA for surveying transcriptional regulation of tumor phenotypes and its potential diagnostic applications in cancer precision medicine.

<u>Experimental Model and Subject Details</u>

*PDX mouse models*

LuCaP patient-derived xenograft tumors (established at the University of Washington) were initiated from tumor specimens resected from men with advanced prostate cancer. The establishment and characterization of the PDX models were described previously (Lam et al. (2018). Generation of Prostate Cancer Patient-Derived

Xenografts to Investigate Mechanisms of Novel Treatments and Treatment Resistance. In Prostate Cancer: Methods and Protocols, Z. Culig, ed. (New York, NY: Springer), pp. 1–27). PDXs were propagated in vivo in male NOD scid IL2R-gamma-null (NSG) mice from Jackson Labs (cat#005557). The collection of tumors for the establishment of PDX lines was approved by the University of Washington Human Subjects Division IRB (IRB #2341) after receiving the patients' written consent. A maximum of five mice were caged in a pathogen-free facility and given unlimited access to food and water maintained on a 12-hour light/dark cycle. Surgeries were performed under isoflurane anesthesia, and mice were given supplemental buprenorphine sustained release (SR). PDX lines were evaluated using histopathology by at least two expert pathologists, and histological phenotypic subtype annotations were orthogonally validated based on transcriptome-derived signature marker expression scores to define phenotypes (Beltran et al. (2016). Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. Nature Medicine *22*, 298–305; Bluemn (2017). Androgen Receptor Pathway-Independent Prostate Cancer Is Sustained through FGF Signaling. Cancer Cell *32*, 474-489.e6; Nyquist et al. (2020). Combined TP53 and RB1 Loss Promotes Prostate Cancer Resistance to a Spectrum of Therapeutics and Confers Vulnerability to Replication Stress. Cell Reports *31*, 107669): adenocarcinoma AR-positive (ARPC), neuroendocrine positive (NEPC), and AR-low, neuroendocrine negative (ARLPC). During LuCaP propagation, if tumors started to ulcerate or if the health of the animal was compromised, the animal was sacrificed and excluded from the study. Resected PDX tumors (300-800 mm$^3$) were divided into ~50mg to ~100mg pieces and stored at -80C. Animal studies were approved by the Fred Hutchinson Cancer Research Center (FHCRC) IACUC (protocol 1618) and performed in accordance with the NIH guidelines. For the current study, blood was collected by cardiac puncture from animals bearing PDX tumors (measurable size 300-800 mm$^3$).

*Human subjects*

UW cohort: Blood samples were collected from men with metastatic castration resistant prostate cancer at the University of Washington (collected under University of Washington Human Subjects Division IRB protocol number CC6932 between years 2014-2021). In this study, 61 plasma samples from 30 patients were analyzed. After initial ultra-low pass whole genome sequencing (ULP-WGS) analysis, 47 plasma samples from 30 patients were retained for further high depth of coverage whole genome sequencing (WGS) analysis. All samples were de-identified prior to ctDNA analysis and a double blinded approach was employed for evaluating clinical phenotype predictions. The initial patient selection was done based on clinical disease burden information and the availability of clinically derived phenotypic subtype annotation. Clinical information on these patients is protected due to IRB protocol restrictions.

DFCI cohort I: Plasma was collected from men diagnosed with mCRPC and treated at the Dana-Farber Cancer Institute (DFCI), Brigham and Women's Hospital, or Weill Cornell Medicine (WCM) between April 2003 and August 2021. All patients provided written informed consent for research participation and genomic analysis of their biospecimen and blood. The use of samples was approved by the DFCI IRB (#01-045 and 09-171) and WCM (1305013903) IRBs. ULP-WGS data at mean coverage 0.5x (range 0.3x – 0.9x) for 101 patients were published previously (Berchuck et al. (2022). Detecting Neuroendocrine Prostate Cancer Through Tissue-Informed Cell-Free DNA Methylation Analysis. Clinical Cancer Research *28*, 928–938). The presence of high-grade neuroendocrine carcinoma of prostate origin was confirmed by two genitourinary pathologists according to modern conventions based on histologic review of available material, reinterpretation of original reports, and integration of available molecular results (Epstein et al. (2014). Proposed Morphologic Classification of Prostate Cancer With Neuroendocrine Differentiation. The American Journal of Surgical Pathology *38*, 756–767). Patients with ARPC (clinically annotated as PRAD) had castration-resistant prostate adenocarcinoma with no pathologic evidence of neuroendocrine differentiation throughout their disease course.

DFCI cohort II: Plasma samples in this cohort were collected from men diagnosed with mCRPC and treated at the Dana-Farber Cancer Institute (DFCI). All patients provided written informed consent for blood collection and the analysis of their clinical and genetic data for research purposes (DFCI Protocol # 01-045 and 11-104). WGS data at mean coverage 27x (range 11x – 44x) (Viswanathan et al. (2018). Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. Cell *174*, 433-447.e19), and ULP-WGS data at mean coverage 0.13x (range 0.07x – 0.18x) (Adalsteinsson et al. (2017). Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nature Communications *8*; Choudhury et al. (2018). Tumor fraction in cell-free DNA as a biomarker in prostate cancer. JCI Insight *3*) were downloaded from dbGAP accession phs001417. Eleven samples from six patients had matching WGS and ULP-WGS with paired-end reads, necessary for analysis by Griffin. Prostate specific antigen (PSA, ng/mL) values and treatment at the time of the blood draw were previously published (Choudhury et al. (2018). JCI Insight *3*). The six patients were treated for adenocarcinoma using Abiraterone, Enzalutamide, or Bicalutamide, or the patients had detectable levels of PSA.

Healthy donor plasmacfDNA WGS data used in this study were obtained from previously published studies. Two samples (HD45 and HD46) with coverage of 13x and 15x, respectively, were accessed from dbGAP under accession phs001417 (Adalsteinsson et al. (2017). Nature Communications *8*; Viswanathan et al. (2018). Cell *174*, 433-447.e19). These donors were consented under DFCI protocol IRB (# 03-022). Thirteen healthy donor plasma cfDNA WGS data (12 male: NPH002, 03, 06, 07, 12, 18, 23, 26, 33, 34, 35, 36; 1 female (used in admixtures): NPH004) with coverages between 13.5x – 27.6x were obtained from the European Phenome Archive (EGA) under accession EGAD00001005343 (Ulz et al. (2019). Nature Communications *10*, 4666).

Method Details

*PDX plasma processing*

Blood samples were collected from NSG mice bearing subcutaneous PDX tumors at the time of sacrifice. The PDX lines were maintained at vivaria in the University of Washington and FHCRC. The blood was processed following methods described for human plasma DNA processing for subsequent DNA isolation. Blood was collected in purple cap EDTA tubes and processed within 4 hours. All blood samples were double spun using centrifugation at 2500g for 10 minutes followed by a 16000g spin of the plasma fraction for 10 minutes at room temperature. For each PDX line, 7-10 mouse plasma samples were pooled. Processed plasma samples were preserved in clean, screw-capped cryo-microfuge tubes and stored at -80°C prior to cfDNA isolation.

*Cell-free DNA isolation*

The QIAamp Circulating Nucleic Acid Kit was used to isolate cfDNA from PDX mouse-derived plasma using the recommended protocol. The pooled plasma samples from 7-10 mice for each PDX line contained ~2-3 mL total plasma volume for each line. The filter retention-basedcfDNA kit method does not implement any fragment size class enrichment. Carrier RNA spike-in was excluded from elusion buffer. Isolated cfDNA was quantified using the Qubit dsDNA HS assay (Invitrogen) and the cfDNA fragment size profiles were analyzed using Tapestation HS D5000 and HS D1000 assays (Agilent).

*Cell-free DNA library preparation and sequencing*

For LuCaP PDX mouse plasma samples, NGS libraries were prepared with 50ng inputcfDNA. Illumina NGS sequencing libraries were prepared with the KAPA hyperprep kit, adopting nine cycles of amplification, and purified using lab standardized SPRI beads. KAPA UDI dual indexed library adapters were used. Library concentrations were balanced and pooled for multiplexing and sequenced using the Illumina HiSeq 2500 at the Fred Hutch Genomics Shared Resources (200 cycles) and Illumina NovaSeq platform at the Broad Institute Genomics Platform Walkup-Seq Services using S4 flow cells (300 cycles). To match with Illumina HiSeq 2500 data, truncated 200 cycles FASTQ files were generated (100 bp paired end reads).

Clinical patient plasma samples collected at University of Washington (UW cohort) were submitted to the Broad Institute Blood Biopsy Services. Briefly, cfDNA was extracted from 2 mL double-spun plasma and ultra-low-pass whole genome sequencing (ULP-WGS) to approximately 0.2x coverage was performed. The ichorCNA pipeline was used to estimate tumor DNA content (i.e., tumor fraction, see below). Forty-seven samples (from 30 patients) had either ≥ 5% tumor fraction or ≥ 2% tumor fraction with AR amplification observed in ichorCNA and were subsequently sequenced to deeper WGS coverage (~20x).

*Cell-free DNA sequencing analysis and mouse subtraction*

AllcfDNA sequencing data used in this study were realigned to the hg38 human reference genome (hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/ hg38.fa.gz). FASTQ files were realigned using BWA mem (Li (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv:1303.3997 [q-Bio]) and post alignment processing was performed according to GATK Best Practices workflow (DePristo et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet *43*, 491–498).

For PDX ctDNA whole-genome sequence data, mouse genome subtraction was performed following the protocol described previously (Jo et al. (2019). Impact of mouse contamination in genomic profiling of patient-derived models and best practice for robust analysis. Genome Biology *20*, 231), wherein reads were aligned using BWA mem to a concatenated reference consisting of both human (hg38) and mouse (mm10, GRCm38.p6, igenomes.illumina.com.s3-website-us-east-1.amazonaws.com/Mus_musculus/NCBI/GRCm38/Mus_musculus_NCBI_GRCm38.tar.gz) reference genomes. Read pairs where both reads aligned to the human reference genome were retained and all other read pairs were removed. Then, remaining reads were re-aligned to the human-only reference. Finally, the GATK best practices workflow was applied to each sample. Following mouse subtraction samples with < 3X depth were removed for downstream analysis. The mouse subtraction pipeline used in this study can be accessed atgithub.com/GavinHaLab/PDX_mouseSubtraction.

*Differential mRNA expression analysis*

RNA isolation of 102 tumors from 46 LuCaP PDX samples was performed as described previously (Labrecque (2019). Molecular profiling stratifies diverse phenotypes of treatment-refractory metastatic castration-resistant prostate cancer. J Clin Invest *129*, 4492–4505). RNA concentration, purity, and integrity was assessed by NanoDrop (Thermo Fisher Scientific Inc) and Agilent TapeStation and RNA RIN >=8 was retained for library preparation. RNA-Seq libraries were constructed from 1 ug of total RNA using the Illumina TruSeq Stranded mRNA LT Sample Prep Kit according to the manufacturer's protocol. Barcoded libraries were pooled and sequenced by Illumina NovaSeq 6000 or Illumina HiSeq 2500 generating 50 bp paired end reads. Sequencing reads were mapped to the hg38 human genome and mm10 mouse genomes using STAR.v2.7.3a (Dobin et al. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21). All subsequent analyses were performed in R-4.1.0. Sequences aligning to the mouse genome and therefor derived from potential contamination with mouse tissue were removed from the analysis using XenofilteR (v1.6) (Kluin (2018). XenofilteR: computational deconvolution of mouse and human reads in tumor xenograft sequence data. BMC Bioinformatics *19*, 366). Gene level abundance was quantitated using the R package GenomicAlignments summarizeOverlaps function using mode=IntersectionStrict, restricted to primary aligned reads. refSeq gene annotations were used for transcriptome analysis. Transcript abundances were input to edgeR (Robinson et al. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140), filtered for a minimum expression level using the filterByExpr function with default parameters, and then limma voom was used for differential expression analysis of NEPC vs. ARPC and ARLPC vs. ARPC. The results were then filtered using a list of 1,635 human transcription factors published previously (Lambert et al. (2018). The Human Transcription Factors. Cell *172*, 650–665), which resulted in 514 genes with FDR<0.05 and fold change > 3. Out of these 514, deregulation of gene expression for 404 transcription factor genes delineated ARPC from NEPC.

*Cleavage Under Targets & Release Using Nuclease (CUT&RUN)*

CUT&RUN is an antibody targeted enzyme tethering chromatin profiling assay in which controlled cleavage by micrococcal nuclease releases specific protein-DNA complexes into the supernatant for paired-end DNA sequencing analysis. CUT&RUN assays were performed for three histone modifications, H3K27ac, H3K4me1, and H3K27me3, according to published protocols (Skene and Henikoff (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. ELife *6*, e21856). CUT&RUN were performed on LuCaP PDX tumors using ~75mg flash-frozen tissue pieces. Briefly, frozen tissues were thoroughly chopped into small pieces and converted into smaller clusters of cells using collagenase and dispase. Cell clusters were made permeabilized using digitonin and nutated with target antibody in EDTA antibody buffer. Time-sensitive micrococcal nuclease enzyme treatments were performed on ice. Released DNA was precipitated along with glycogen career, and subsequent NGS libraries were prepared using picogram input DNA library preparation protocol.

Paired-end (50 bp) sequencing was performed and reads were aligned using bowtie2 version 2.4.2 (Langmead et al. (2019). Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics *35*, 421–432) to the hg38 human reference assembly. Aligned reads were processed as described in the SEACR protocol (github.com/FredHutch/SEACR#preparing-input-bedgraph-files). Peaks were called using SEACR version 1.3 (Meers et al. (2019). Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. Epigenetics & Chromatin *12*, 42) using "stringent" settings and with reference to paired IgG controls. Bam files were viewed, parsed, and filtered using SAMtools (Danecek et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008). Insert size restricted analysis was performed by retaining <120 bp fragments (sub-nucleosomal sized), or retaining 140-200 bp fragments (nucleosomal sized). BigWig files were prepared using bamCoverage in deepTools 3.5.0 (Ramírez et al. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research *44*, W160–W165) with a bin size of 10 and the "extendReads" option on bam files. Genomewide peak heatmap, targeted heatmap, and respective profiles were plotted using deepTools. bigWig formatted files for each phenotype were obtained using the mean function in wiggletools 1.2.8. and deepTools computeMatrix. Phenotype-specific informative region coordinates were obtained from diffBind v3.5.0, and the top 10,000 most significant regions (all with FDR < 0.05) differentially open between ARPC and NEPC lines were used for downstream feature analyses (see Gene body and promoter region selection for additional subsetting criteria applied on a feature by feature basis). For heatmaps and profiles the plotHeatmap function was used. The "Peak Center" option was used to derive desired heatmaps. These steps were all performed for H3K27ac, H3K4me1 and H3K27me3 antibodies. Scaled heatmap profiles' area under the curve (AUC) and peak height at the profile center were estimated using deepStats v0.4 (Richard, 2020) (comparable profiles are scaled to 10 units).

*Differential histone post-translational modification (PTM) analysis*

Differential PTM analysis was performed with the Diffbind version 2.16.0 package (Ross-Innes et al. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature *481*, 389–393) in R-4.0.1 using standard parameters (bioconductor.riken.jp/packages/3.0/bioc/html/DiffBind.html). ARPC, NEPC and ARLPC samples were grouped by histopathological and transcriptome signature defined phenotypes described in the "PDX mouse models" section. Samples were loaded with the dba function, reads counted with the dba.count function, and contrast specified as phenotype with dba.contrast and a minimum members of 2. Differential peak sites were computed with the dba.analyze function with default settings. Differential peak binding of NEPC and ARLPC was computed against ARPC samples. Unique binding sites in NEPC and ARLPC were catalogued using bedtools v2.29.2 (Quinlan and Hall (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842). Intergroup differentially bound peaks were annotated using ChIPseeker 1.28.3 (Yu et al. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics *31*, 2382–2383) and TxDb.Hsapiens.UCSC.hg38.knownGene 3.2.2 in R 4.1.0.

*ATAC-Seq analysis*

ATAC-Seq sequence data for 15 tumor samples from 10 PDX lines were published previously and FASTQ files made available upon request (Cejas et al. (2021). Subtype heterogeneity and epigenetic convergence in neuroendocrine prostate cancer. Nat

Commun *12*, 5775). These lines included LuCaP PDX lines with ARPC histology (23.1, 77, 78, 81, 96) and NEPC histology (two replicates each of 49, 93, 145.1, 173.1 and one replicate of 145.2). Paired end reads were aligned using bowtie2 2.4.2 (Langmead et al. (2019). Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics *35*, 421–432) aligned to the UCSC hg38 human reference assembly with the "very-sensitive" "-k 10" settings. Peaks were called using Genrich version 0.6.1 (github.com/jsh58/Genrich). Differential binding analysis was performed using Diffbind version 3.5.0 package in R version 4.1.0. ENCODE blacklisted regions were excluded using hg38-blacklist.v2 (Amemiya et al. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep *9*, 9354) (github.com/Boyle-Lab/Blacklist). AR positive, NE null PDX samples (n = 5) were compared against AR null, NE positive PDX samples (n = 5) using RNA-Seq derived phenotypes. Phenotype specific binding sites were isolated by first selecting for positive fold change open chromatin enrichment and then using Intervene 0.6.5 (Khan and Mathelier (2017). Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. BMC Bioinformatics *18*, 287) where regions were considered overlapping if they shared at least 1 bp. Regions with FDR adjusted p-values < 0.05 were then subset to those overlapping the 338,000 established TFBSs (338 TFs x 1,000 binding sites, see Griffin analysis for site selection) by at least 1 bp using BedTools Intersect. Only regions that overlapped an established TFBS were retained.

*Griffin analysis*

Griffin is a method for profiling nucleosome protection and accessibility on predefined genomic loci (see Example 1 and Doebley et al. (2021). Griffin: Framework for clinical cancer subtyping from nucleosome profiling of cell-free DNA. MedRxiv 2021.08.31.21262867). Griffin filters sites by mappability, estimates and corrects GC bias on a per fragment level, and generates GC-corrected coverage profiles around each site. First, griffin takes a site list and examines the mappability in a window (+/- 5000 bp around each site). Mappability (hg38 Umap multi-read mappability for 50bp reads) was obtained from UCSC genome browser (Karimzadeh et al. (2018). Umap and Bismap: quantifying genome and methylome mappability. Nucleic Acids Research *46*, e120) (hgdownload.soe.ucsc.edu/gbdb/hg38/hoffmanMappability/k50.Umap.MultiTrackMappability.bw). Sites with <0.95 mappability were excluded from further analysis. Next, GC bias was quantified for each sample using a modified version of the approach described previously (Benjamini and Speed (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Research *40*, e72–e72). Briefly, for each possible fragment length and GC content, the number of reads in a bam file and the number of genomic positions with that specific length and GC content were counted. The GC bias for each fragment length and GC content was calculated by dividing the number of observed reads by the number of observed genomic positions for that fragment length and GC content. The GC bias for all possible GC contents at a given fragment length was then normalized to a mean bias of 1. GC biases were then smoothed by taking the median of values for fragments with similar lengths and GC contents (k nearest neighbors smoothing) to generate smoothed GC bias values.

After GC correction, nucleosome profiling was performed in each sample. For each mappable site of interest, fragments aligning to the region ± 5000 bp from the site were fetched from the bam file. Fragments were filtered to remove duplicates and low-quality alignments (<20 mapping quality) and by fragment length. Nucleosome size fragments (140-250 bp) were retained. Fragments were then GC corrected by assigning each fragment a weight of 1/GC_bias for that given fragment length and GC content and the fragment midpoint was identified. The number of weighted fragment midpoints in 15bp bins across the site were counted. For composite sites, all sites of a given type (such as all sites for a given transcription factor) were summed together to generate a single coverage profile. Individual or composite coverage profiles were normalized to a mean coverage of 1 in the ± 5000bp region surrounding the site. Finally, sites were smoothed using a Savitsky-Golay filter with a window length of 165bp and a polynomial order of 3. The window ± 1000 bp around the site was retained for plotting and feature extraction (See Griffin manuscript for further details); when plotting sites, shading illustrates the 95% confidence interval within sample groups. Features extracted from individual or composite sites included:

a. "mean central coverage," the mean coverage between -30 to 30 bp relative to the site center,
b. "mean window coverage," the mean coverage between -990 to 990 bp relative to the site center, and
c. "max wave height," the absolute difference between the minimum coverage within the window from -120 to 30 bp and maximum coverage in the window from 31 to 195 bp relative to the TSS.

*Griffin analysis for selective transcription factor binding sites (TFBS)*

Transcription factor binding site (TFBS) Griffin analysis was conducted with the same TFBS list utilized in Griffin (see Example 1 and Doebley et al. (2021). MedRxiv 2021.08.31.21262867). Briefly, TFBS locations were downloaded from the Gene Transcriptional Regulation Database (GTRD) (Yevshin et al. (2019). GTRD: a database on gene transcription regulation—2019 update. Nucleic Acids Res *47*, D100–D105), which contains a compilation of ChIP-Seq data from multiple sources. The meta cluster data or meta peaks (version 19.10) observed in one or more ChIP-Seq experiments was used. The initial list of 1,314 transcription factors (TFs) in GTRD was compared to the CIS-BP database (Weirauch et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–1443) (v2.00 downloaded from cisbp.ccbr.utoronto.ca/bulk.php). TFs from GTRD that were also in CIS-BP and had a known binding motif were retained. Selected TF binding genomic loci were then filtered for mappability as described above (Griffin analysis) and TFs with fewer than 10,000 highly mappable sites on autosomes were excluded, resulting in 338 TFs. For each TF, the sites were sorted by the 'peak.count' value from GTRD and used the 1,000 sites (genomic loci) with the highest number of peaks for downstream analysis by Griffin. After intersecting these 338 with 404 differentially expressed TFs identified through RNA-Seq 107 remained, on which unsupervised hierarchical clustering of central window mean values was performed (see Griffin analysis). Hierarchical clustering was performed using the Ward.D2 method with Euclidean distance and complete linkage settings; the groupings were determined using cutree_cols=2 for columns (LuCaP CRPC phenotypes) and cutree_rows=13 for rows (TFs) on the dendrograms.

*Gene body and promoter region selection*

For individual gene body and promoter analyses Ensembl BioMart v104 (hg38) (Howe et al. (2021). Ensembl 2021. Nucleic Acids Research *49*, D884–D891) was used to directly retrieve protein coding transcript start (TSS) and end (TES) coordinates. For promoter region analysis the window ±1000 bp relative to the TSS was considered. For gene body analysis, the region between the TSS and TES was considered. In the case of genes with multiple transcripts, analyses were limited to the longest transcript resulting in 19,336 regions. In downstream analysis of LuCaP PDX cfDNA, if any lines did not meet specific criteria in a region (including differentially open histone modification regions) that feature/region combination was excluded from analysis, leading to a variable lower number of regions considered based on the feature. These criteria included requiring at least 10 total fragments in a region for all Fragment size analysis (see below) and a non-zero number of "short" and "long" fragments for the short-long ratio; short-long ratios less than 0.01 or greater than 10.0 were also excluded as outliers. Any region with no coverage in a line was excluded from all analyses. This resulted in gene lists that differed in numbers between genomic contexts and feature types.

*Fragment size analysis*

Fragments were first filtered to remove duplicates and low-quality alignments (<20 mapping quality) and by fragment length (15-500 bp). In individual genomic loci/windows, the fragment short-long ratio (FSLR) was computed as the ratio of short (15 - 120 bp) to long (140 - 250 bp) fragments. The mean, median absolute deviation (MAD: $median(|X_i - median(X)|)$), and coefficient of variation (CV: $\frac{\sigma}{\mu}$ where σ = standard deviation, μ = mean) of the fragment length distribution was also calculated for each selected window. The fragment size analysis code and implementation used in this study can be accessed at github.com/GavinHaLab/CRPCSubtypingPaper/tree/main/FragmentAnalysis.

*ctDNA tumor-normal admixtures and benchmarking*

Admixtures for evaluating benchmarking performance were constructed using 5 ARPC (LuCaP 35, 35CR, 58, 92, 136CR) and 5 NEPC (LuCaP 49, 93, 145.2, 173.1, 208.4) lines mixed to 1%, 5%, 10%, 20%, and 30% tumor fraction with a single healthy donor plasma line (NPH004, EGAD00001005343) at ~25X mean coverage, assuming 100% tumor fraction in post-mouse subtracted PDX sequencing data. After extracting chromosomal DNA with SAMtools (Danecek et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008) and removing duplicates with Picard (broadinstitute.github.io/picard/), SAMtools was used to merge BAM files. Admixtures were then down-sampled to the number of reads corresponding to 1X and 0.2X using SAMtools to evaluate (ultra) low-pass WGS performance. During unsupervised benchmarking of each admixture the healthy and LuCaP line used in the admixture were excluded from the generation of feature distributions to ensure the model would not learn from the lines being interrogated. The admixture pipeline used in this study can be accessed at github.com/GavinHaLab/Admixtures_snakemake.

*Supervised binary classification of ARPC and NEPC*

Binary classification of ARPC and NEPC subtypes using individual region and feature combinations was conducted using XGBoost 'XGBClassifier' implemented in Python with default parameters. Features included histone modification regions, promoters, and gene bodies; fragment size mean, short-long ratio, and coefficient of variation (see Fragment size analysis) in histone modification regions, and promoters; central and window coverage (see Griffin analysis) in promoters, composite TFBSs, and composite differentially open chromatin regions identified through ATAC-Seq; and Max Wave Height (See Griffin analysis) in promoters. Stratified 6-fold cross-validation was applied where two ARPC samples and one NEPC sample was held out in each fold. This was repeated 100 times and performance was computed using area under the receiver operating characteristic (ROC) curve (AUC) and 95% confidence intervals for each individual feature and region combination. Code and implementation of the method can be found at github.com/GavinHaLab/CRPCSubtypingPaper/tree/main/SupervisedLearning.

*ichorCNA tumor fraction estimation*

Tumor fractions from patient plasma samples were assessed using ichorCNA (Adalsteinsson et al. (2017). Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nature Communications *8*) with binSize 1,000,000 bp and hg19 reference genome. Default tumor fraction estimates reported by ichorCNA were used. See github.com/GavinHaLab/CRPCSubtypingPaper/tree/main/ichorCNA_configuration for complete configuration settings.

*Phenotype prediction model*

A probabilistic model was developed to classify the mCRPC phenotype (ARPC or NEPC) in an individual patient plasma ctDNA sample. This is a generative mixture model that is unsupervised—it does not train on the patient cohort of interest. However, the model accepts the pre-estimated tumor fraction from ichorCNA for the given patient ctDNA sample, as well as the pre-computed ctDNA features values from the LuCaP PDX ctDNA and healthy donor ctDNA as prior information. For each patient ctDNA sample, it fits the heterogeneous tumor fractions against the pure PDX LuCaP models. The expected feature value (mean μ and standard deviation σ) from each phenotype $k$ for feature $i$ were taken from the mean of LuCaP PDX samples ($\mu_{i,k}$), or taken from the mean of a panel of normals $H$ ($\mu_{i,H}$, male only, n = 14; see Healthy Donor cohort) assuming a Gaussian distribution, is shifted such that the shifted values $\mu'_{i,k}$ , $\sigma'_{i,k}$ took the form:

$$\mu'_{i,k} = \alpha\mu_{i,k} + (1 - \alpha)\mu_{i,H}$$
$$\sigma'_{i,k} = \sqrt{\alpha\sigma^2_{i,k} + (1 - \alpha)\sigma^2_{i,H}}$$

where $\alpha$ is the tumor fraction estimate for each test sample. In the final model, four features were used: composite open chromatin regions (central and window mean coverage) for specific phenotypes (ARPC and NEPC) identified from the LuCaP PDX ATAC-Seq analysis using Griffin (see Griffin analysis). For each feature *i*, the probability that the observed sample came from a mixture of the tumor-fraction-corrected Gaussian distributions was found, where $\theta$ is the NEPC mixture weight:

$$p_i(x|\theta) = \theta p(x|k = NEPC) + (1 - \theta)p(x \mid k = ARPC)$$

The $\theta$ parameter is estimated by maximizing the joint log-likelihood $L$ for a given patient sample:

$$\theta' = \underset{\theta}{\operatorname{argmax}}[L(x|\ \theta)]$$

$$where\ L(x|\ \theta) = \sum_i \ln\,[p_i(x|\theta)]$$

$\theta$ has range [0,1], where higher values indicate an increased proportion of the sample having a NEPC phenotype and was used as the NEPC prediction score metric. Code and implementation of the method can be found at github.com/GavinHaLab/CRPCSubtypingPaper/tree/main/GenerativeMixtureModel.

*Analysis and classification of clinical patient samples*

After establishing feature distributions using the LuCaP PDX lines and normal panel as described in Generative model, the model was applied to three clinical patient cohorts (see Human subjects for cohort information). Initial scoring using the model was run on DFCI cohort I, consisting of 101 ULP-WGS samples with paired-end reads. Tumor fraction estimates predicted by ichorCNA in the original study (Berchuck et al.

(2022). Detecting Neuroendocrine Prostate Cancer Through Tissue-Informed Cell-Free DNA Methylation Analysis. Clinical Cancer Research *28*, 928–938) and tumor phenotype classifications were obtained from the original study. A prediction score threshold of 0.3314 for calling NEPC was chosen because it offered an optimal performance for sensitivity (90%) and specificity (97.5%), where sensitivity is the true positive rate for identifying NEPC samples $\left(\frac{TP}{TP+FN}\right)$ and specificity is the true negative rate for identifying ARPC samples $\left(\frac{TN}{TN+FP}\right)$. Alternative thresholds maximizing sensitivity and specificity were 0.1077, at which 95% sensitivity was achieved with a lower specificity of 93.8%, and 0.3769 with a lower sensitivity of 81.0% but higher specificity of 98.8%.

The model was then validated on two cohorts, beginning with the already published DFCI cohort II (Adalsteinsson et al. (2017). Nature Communications *8*; Choudhury et al. (2018). Tumor fraction in cell-free DNA as a biomarker in prostate cancer. JCI Insight *3*; Viswanathan et al. (2018). Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. Cell *174*, 433-447.e19). The analysis was restricted to eleven samples from six patients with matched ULP-WGS and WGS data with paired-end reads. Tumor fraction estimates from ichorCNA were obtained from the original study ( Adalsteinsson et al. (2017). Nature Communications *8*). All samples were considered adenocarcinoma (ARPC) based on clinical histories (see Human subjects). The scoring threshold of 0.3314, determined from DFCI cohort I was used for phenotype classification.

For the *UW cohort*, consisting of 47 samples from 30 patients, ichorCNA was used to estimate sample tumor fractions as described above, while clinical phenotype was determined from clinical histories and expert chart review. Model performance was evaluated on matched ULP-WGS and WGS data for unambiguous clinical phenotypes of ARPC and NEPC. The chosen scoring threshold of 0.3314 was used, and the fraction of correctly predicted ARPC (n=26) and NEPC (n=5) was computed. The remaining 16 samples with mixed histologies were not evaluated for performance.

Quantification and Statistical Analysis

Quantification of and statistical approaches for high-throughput sequencing data analysis are described in the methods above. When non-parametric distributions (not normally distributed) of numerical values of a particular parameter in a population were compared (using boxplots or in tables), the two-tailed Mann-Whitney U test (also known as the Wilcoxon Rank Sum test; scipy.stats.mannwhitneyu, (Virtanen et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods *17*, 261–272) was used to test if any two distributions being compared were significantly different, with Benjamini-Hochberg (statsmodels.stats.multitest.fdrcorrection, statsmodels.org) correction applied in multiple testing scenarios. All boxplots represent the median with a centerline, interquartile range (IQR) with a box, and first quartile – 1.5 IQR and third quartile + 1.5 IQR with whiskers. PCA was conducted in Python (sklearn.decomposition.PCA; scikit-learn.org)

## Example 3

Example 1 applied an embodiment of the Griffin workflow to enhance sequence signals to allow accurate determination of breast cancer subtypes from low pass sequencing data. Example 2 applied an embodiment the Griffin workflow approach to differentiate subtypes of other cancers, namely prostate cancer, successfully leveraging data from an alternative sequence profiling platform (e.g., from the CUT & RUN platform for nucleosome accessibility), demonstrating the power and flexibility of the Griffin analytic workflow for different cancers and input data. This Example described the development of targeted sequencing panels to use in conjunction with the Griffin workflow to understand transcriptional features of small cell lung cancer, non-small cell lung cancer, and other cancer types from blood ctDNA.

### Background

This Example describes an innovative analytical assay based on analysis of cell free DNA, demonstrating clear translational potential for clinical lung cancer diagnostics. Cell-free DNA circulating in the blood of cancer patients has been widely used to assess gene mutations, and through analyses of whole genome DNA has more recently been used to infer activation of certain transcription factors. Cancer cells give rise to cell-free DNA via cell death and that cell-free DNA is overwhelmingly nucleosomal, i.e. bound to a histone octamer, which protects the DNA from degradation. Histone positioning in the genome is influenced by components of chromatin, including transcription factors and the RNA polymerase complex. At transcription factor binding sites (TFBSs) that are bound to the cognate transcription factor (TF), or transcription start sites (TSSs), corresponding to highly expressed genes, nucleosomes are displaced. Sophisticated analysis of cell-free DNA sequencing data can therefore reveal nucleosome location, and in turn, TF occupancy and gene expression, in the cells of origin. However, analyses performed thus far have largely used deep whole genome sequencing of ctDNA, which is not practical and cost effective as a clinical test to be widely applied to lung cancer patients.

Results and Discussion

One innovation is the identification of highly informative TFBSs and TSSs that can be used to differentiate between NSCLC and SCLC or between subtypes of NSCLC or SCLC, which then facilitate the use of hybridization capture-based DNA sequencing of ctDNA to generate high resolution maps of nucleosome occupancy at TFBSs of key TFs in SCLC (ASCL1, NEUROD1, POU2F3, REST) and TSSs for genes that are markers of key transcriptional features of lung cancer cells. Alternatively, these informative sites can also be examined in low-coverage whole genome sequencing to extract similar transcriptional features. Targeted capture panels are routinely applied in the clinic to call mutations from ctDNA in the blood, and application of targeted sequencing to assess transcriptional activity in cancer cells is very feasible as a clinical test. The technology is especially relevant and viable in SCLC, which kills ~30,000 people in the US each year. Tissue sampling in SCLC is typically only performed once during a patient's disease course and is often done by transbronchial fine needle aspiration, which yields a very small amount of tissue. Surgery is very rarely performed. However, SCLC has a high level of ctDNA compared to most other cancer types, reflecting its highly metastatic nature, making this assay both practical for application to SCLC patients and potentially especially valuable. Also, there is increasing realization that SCLC subtypes exist based not on mutations but on activation of key transcription factors and their downstream programs (such as ASLC1, NEUROD1, and POU2F3). Despite clinical urgency to characterize these subtypes, there is no established technique for determining transcriptional subtype in SCLC using blood samples. The disclosed targeted assay is designed to differentiate transcriptional subtypes of SCLC from ctDNA providing powerful clinical applications for use of this assay. Additionally, the panel is designed to call gene mutations in exons from a panel of ~600 genes. Thus, the assay has broad clinical utility for correlative analyses of both mutations and transcriptional activity in clinical samples.

Another significant current challenge in lung cancer management is transdifferentiation of driver mutation positive lung cancer to SCLC, which typically occurs after an extended period of disease control with targeted therapy.

Transdifferentiation to SCLC is treated differently from disease that is progressing but has not acquired a notable histologic change. However, transdifferentiation is likely significantly underdiagnosed because currently it can only be assessed via biopsy of a progressing lesion, which is often infeasible or undesirable. This assay can also be applied to lung adenocarcinoma patients who develop resistance to EGFR inhibitors to determine whether this resistance is associated with activation of SCLC transcriptional profiles. Thus, the major non-invasive applications include the following:

- tumor classification of SCLC by estimating the activity of transcription factors (ASCL1, NEUROD1, POU2F3, REST)
- distinguish SCLC from NSCLC
- distinguish between the major NSCLC histological subtypes: adenocarcinoma and squamous cell carcinoma
- estimate mixed histologies
- detect potential subtype changes during therapy in "real-time" (either treatment-induced transcriptional subtype changes within SCLC, or change from NSCLC to SCLC, which occurs in NSCLC as a resistance mechanism towards targeted therapies such as EGFR inhibition).

All of these applications have critical implications for studying novel therapies and informing clinical treatment decisions.

A schematic overview of the panel design is provided in FIG. 16, which shows the generation of a capture panel. In more detail, the approach included rationally designing a targeted sequencing panel for integrated detection of SCLC genetic mutations, transcription factor (TF) subtype identity, and expression of key gene programs. Public mutations databases and functional mutation data were interrogated for coding mutations coding in approximately 600 genes related to SCLC. For TF subtype identity, TFBSs for four key SCLC-related TFs (ASCL1, NEUROD1, POU2F3, and REST) were targeted. For expression of other key gene programs, TSSs corresponding to the vast majority of protein-coding genes in the genome were targeted. To select specific sites, multiple sources of data were integrated as follows. For the SCLC TFBSs, ChIP-seq data was used to identify TFBSs, resulting in 4-30k sites per factor. These candidate sites were then annotated with the distance to the nearest gene TSS. Retained sites were sites for which the nearest gene TSS was a gene known to be upregulated in SCLC cells that expressed the factor of interest, as determined by available RNAseq data. This resulted in ~400-700 SCLC-focused sites per factor. In the final probe set, a 1 kb window symmetrically encompassing these ~2k sites (500 bp on each side) was targeted. For TSS profiling, beginning with an established transcript annotation, non-coding transcripts, Y chromosome genes, and TSSs corresponding to multi-exon genes that had lower confidence annotations were removed, resulting in approximately ~36k theoretically targeted TSSs. In the probe set, regions 260 bp downstream of the TSS and 100 bp upstream were targeted. Use of application-specific orthogonal chromatin profiling data to select sites is a key feature of the approach. However, it will be noted that other types of chromatin profiling data could readily be substituted or added and yield same or similar results, such as ATAC-seq, CUT&RUN/TAG, DNAse-seq, modified histone ChIP-seq, etc.

A data analysis pipeline was developed to quantify cfDNA fragments protected by nucleosomes in both the TFBS and TSS captured DNA. The analysis pipeline, Griffin (described in more detail above), includes using fragment length-based GC correction to remove GC biases that obscure signals. A fragment size-aware GC-bias correction approach helps to maximize signal-to-noise and optimizes the analysis of captured DNA.

FIGS. 17A and 17B illustrate the detection of transcription factor (TF) expression in SCLC models using targeted sequencing of cfDNA. FIG. 17A is a schematic of experimental workflow for proof-of-concept negative control ("healthy donor") and positive control ("flank tumors" from SCLC cellular models) samples. FIG. 17B graphically illustrates aggregated coverage across TFBSs in targeted sequencing data for healthy donors (top row) and flank tumors (bottom row). The TFBS is expected to be located at position 0 on the x axis. Data are coded by expected TF expression. Healthy donor-derived cfDNA is expected to reflect REST expression but not ASCL1, NEUROD1, or POU2F3. In SCLC models, systematic differences in coverage distribution as a function of TF expression are apparent.

FIGS. 18A-18C illustrate transcription factor activity inference using TFBS coverage distributions from SCLC patient samples with available matched tumor gene expression data. FIG. 18A graphically illustrates aggregated coverage across TFBSs in targeted sequencing data for healthy donors (top row) and patients with SCLC (bottom row) for whom matched tumor tissue with gene expression data was available. Samples are coded by expected TF expression. Systematic differences in coverage distribution as a function of expected TF expression are again apparent. FIG. 18B illustrates gene expression of key genes in selected patient samples displayed as a heatmap. Cells are coded by Z-score and the inset text is the log2(TPM+1). FIG. 18C illustrates peak to trough amplitude calculated from coverage distributions at TFBS in each patient sample displayed as a heatmap. The amplitude is displayed by color and also as inset text. Trough depth magnitude corresponds to gene expression of the key TFs in these bona fide SCLC patient samples.

FIG. 19 is a series of graphs illustrating quantification of transcription factor binding site peak to trough amplitude sample types. Distribution of TFBS peak to trough amplitude calculated from aggregated coverage distributions according to expected ground truth of TF expression. Pdx samples labeled "not SCLC" are NSCLC pdx models. Patient samples labeled "not SCLC" are either samples from patients with NSCLC (n=11) or without a diagnosis of malignancy (n=4). ASCL1 site peak to trough amplitude is associated with both SCLC status and ASCL1 positivity, while NEUROD1 and POU2F3 peak to trough amplitude is associated only with TF positivity.

FIGS. 20A and 20B graphically illustrate gene expression inference using TSS coverage distributions in flank tumor positive control samples. FIG. 20A illustrates TSS coverage distribution from targeted sequencing of cfDNA, grouped by gene expression quintile in SCLC flank tumor models (quintiles 1-5) and blood ("B", dark blue). Shown are 1,912 TSS corresponding to 1,213 genes, which were selected based on low expression in whole blood and correlation between TSS coverage distribution and gene expression. TSS coverage distribution varies systematically according to expression of the corresponding gene. FIG. 20B illustrates receiver operating characteristic curves for prediction of gene expression as above or below a threshold value (shown for thresholds of 0.1, 0.5, 1.0, and 2.0), as inferred from the coverage distribution of the corresponding TSS. An estimator of gene expression was calculated from the TSS coverage profile as the magnitude of the difference of the average coverage depth at positions +130 and +145 relative to the TSS minus the average depth at positions -45, -30, and -15 (shown as a dotted line in 20A). The AUC of the ROC curve is shown in parentheses for each gene expression cutoff. TSS coverage distributions can be used to predict whether a gene is expressed above or below a certain value with good test characteristics in this preliminary analysis that is restricted to especially variable, and therefore challenging, genes.

FIG. 21 is a series of graphs illustrating use of aggregated coverage profiles across large rationally selected subsets of the TSS panel for prediction of SCLC vs NSCLC status in lung cancer Pdx models and Patient samples. The graphs provide examples of aggregated TSS coverage distributions across gene TSSs selected for upregulation in NSCLC (n=396) and SCLC (n=1045) for three different samples: one healthy donor, one NSCLC Pdx model, and one SCLC Pdx model. As shown overlayed on the NSCLC PDX model, an amplitude feature was calculated from each coverage distribution curve as the difference between the coverage at the -45 position and the +120 position relative to the TSS, facilitating comparison within and between samples.

FIG. 22 is a series of graphs illustrating use of aggregated coverage profiles across large rationally selected subsets of the TSS panel for prediction of SCLC vs NSCLC status in lung cancer Pdx models and Patient samples. Aggregate coverage of SCLC-specific gene TSS (y axis, n=1045) vs NSCLC-specific gene TSS (x axis, n=396) in plasma samples from lung cancer PDX samples (non-cancer control patients also shown for reference as "benign") or from lung cancer patients. An SCLC PDX that transdifferentiated from an adenocarcinoma is identified with a thick red line.

Accordingly, these data show that the rationally designed capture panel allowed harvesting of sequence data that, when optimized using the Griffin workflow and with application of appropriate classifiers, can accurately differentiate SCLC and NSCLC cells from both PDF models and patient samples. There are clear technical advantages throughout the disclosed workflow.

Griffin uses unique normalization of cfDNA sequence data that is specific for nucleosome profiling and chromatin accessibility analysis. This includes GC-bias correction, repetitive sequence filtering, and local coverage normalization. All of these normalization techniques are not available in existing proof-of-concept methods such as in Ulz P, et al. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nat Commun. 2019;10(1):4666. Further, multi-omic feature extraction from Griffin for use in machine learning classifier construction to predict cancer subtype is unique to this approach. Especially for resolution of more similar cell types, use of a targeted sequencing panel is expected to yield higher resolution while retaining practical cost, and is more readily integrable with resequencing of regions of interest for genetic mutation detection (i.e. cancer gene panel sequencing). From output of Griffin, many features can be extracted from each binding site of interest and machine learning classifiers can be used to predict subtypes of lung cancer histological subtypes from the cfDNA Griffin-optimized data.

While illustrative embodiments have been illustrated and described, it will be appreciated that various changes can be made therein without departing from the spirit and scope of the invention.

# CLAIMS

The embodiments of the invention in which an exclusive property or privilege is claimed are defined as follows:

1. A computer-implemented method of enhancing sequence read data from cell-free DNA samples for cell type prediction, the method comprising:

receiving, by a computing system, sequence read data, wherein the sequence read data includes a plurality of fragment reads, wherein each fragment read has a fragment length and a GC content indicating a percentage of bases in the fragment read that are G or C;

determining, by the computing system, GC bias values for each fragment read based on the fragment length and the GC content of the fragment read;

generating, by the computing system, a genomic coverage distribution that is adjusted for GC bias using the sequence read data and the GC bias values; and

predicting, by the computing system, the cell type based on the genomic coverage distribution.

2. The computer-implemented method of claim 1, wherein predicting the cell type based on the genomic coverage distribution includes predicting a cell phenotype.

3. The computer-implemented method of claim 2, wherein predicting the cell phenotype includes predicting a tissue type, a cancer type, or a cancer subtype.

4. The computer-implemented method of claim 2, wherein predicting the cell phenotype includes predicting expression of one or more genes of interest.

5. The computer-implemented method of claim 1, wherein determining the GC bias value based on the fragment length and the GC content of the fragment read includes:

counting a number of observed reads of each combination of fragment length and GC content to determine GC counts for the sequence read data;

dividing the GC counts by corresponding GC frequencies in a GC frequency matrix to determine a GC bias for each fragment length;

normalizing a mean GC bias for each fragment length to determine rough GC bias values; and smoothing the rough GC bias values to determine the GC bias values.

6. The computer-implemented method of claim 5, wherein the GC frequency matrix stores a frequency for each GC content for each fragment length of a plurality of fragment lengths in mappable regions of a reference genome.

7. The computer-implemented method of claim 6, wherein the plurality of fragment lengths includes each fragment length from a short length threshold to a long length threshold.

8. The computer-implemented method of claim 7, wherein the short length threshold is in a range of 10-20 base pairs, and wherein the long length threshold is in a range of 450-550 base pairs.

9. The computer-implemented method of claim 8, wherein the short length threshold is 15 base pairs, and wherein the long length threshold is 500 base pairs.

10. The computer-implemented method of claim 1, further comprising:

determining genomic regions of interest for a cell type; and

filtering the genomic regions of interest to identify cell-type-informative sites.

11. The computer-implemented method of claim 10, wherein determining the genomic regions of interest includes:

determining a mean mappability in a fixed size window around each genomic region of interest; and

discarding genomic regions of interest having a mean mappability less than a predetermined threshold.

12. The computer-implemented method of claim 10, wherein filtering the genomic regions of interest to identify cell-type-informative sites includes determining sites that have differential signals between a first cell type and a second cell type.

13. The computer-implemented method of claim 10, wherein generating the genomic coverage distribution includes:

determining fragment midpoints in a window around each cell-type-informative site;

assigning a weight for each fragment read based on an inverse of the GC bias value for each fragment read;

using the weighted fragment reads to determine GC-corrected midpoint coverage profiles;

excluding positions that overlap excluded regions;

determining a mean profile based on determining an average of GC-corrected midpoint coverage profiles for all sites;

smoothing the mean profile to generate a smoothed mean profile; and

normalizing the smoothed mean profile by dividing by a mean of surrounding coverage to determine a normalized mean profile.

14. The computer-implemented method of claim 13, wherein the excluded regions include one or more regions that are within an encode unified GRCh38 exclusion list, centromeres, gaps in human genome assembly, fix patches, alternative haplotypes, regions of zero mappability, or have coverage of at least 10 standard deviations above a mean.

15. The computer-implemented method of claim 10, wherein predicting the cell type based on the genomic coverage distribution includes:

generating one or more features based on the genomic coverage distribution;

providing the one or more features as input to a classifier model; and

determining the cell type based on an output of the classifier model.

16. The computer-implemented method of claim 15, wherein the one or more features include a mean of coverage in a first predetermined window around each cell-type-informative site, a mean of coverage in a second predetermined window of a different size than the first predetermined window around each cell-type-informative site, and an amplitude of the genomic coverage distribution around each cell-type-informative site.

17. The computer-implemented method of claim 16, wherein the first predetermined window is larger than the second predetermined window.

18. The computer-implemented method of claim 17, wherein the first predetermined window has a width in a range of 1800-2200 base pairs, and wherein the second predetermined window has a width in a range of 40-80 base pairs.

19. The computer-implemented method of claim 18, wherein the first predetermined window has a width of 2000 base pairs, and wherein the second predetermined window has a width of 60 base pairs.

20. The computer-implemented method of claim 16, wherein the amplitude of the genomic coverage distribution around each cell-type-informative site is determined by:

trimming the genomic coverage distribution to a window that contains 10 peaks;

performing a fast Fourier transform on the window of the genomic coverage distribution; and

determining a magnitude of the 10th frequency.

21. The computer-implemented method of claim 15, wherein the classifier model includes a logistic regression model, an artificial neural network, a decision tree, a support vector machine, or a Bayesian network.

22. A method of determining a chromatin accessibility profile for a cell of interest from a sample comprising cell-free DNA derived from the cell of interest, the method comprising:

obtaining sequence read data from the cell-free DNA;

receiving, by a computing system, sequence read data, wherein the sequence read data includes a plurality of fragment reads, wherein each fragment read has a fragment length and a GC content indicating a percentage of bases in the fragment read that are G or C;

determining, by the computing system, GC bias values for each fragment read based on the fragment length and the GC content of the fragment read;

generating, by the computing system, a genomic coverage distribution that is adjusted for GC bias using the sequence read data and the GC bias values; and

determining the chromatin accessibility profile from the genomic coverage distribution.

23. The method of claim 22, further comprising determining a phenotype of the cell of interest based on the chromatin occupancy profile.

24. The method of claim 23, wherein determining the cell phenotype comprises determining a tissue type, a cancer type, a cancer subtype, a malignancy aggressiveness phenotype, and/or a drug responsivity phenotype.

25. The method of claim 22, further comprising performing one or more steps as recited in one or more of claim 5 to claim 21.

26. A method for determining a cell type of a cell of interest from a sample comprising cell-free DNA derived from the cell of interest, comprising:

obtaining sequence read data generated from the sample comprising cell-free DNA;

performing the computer-implemented method recited in any one of claims 5 to 21; and

determining the cell type of the cell of interest based on the prediction provided by the computing system.

27. The method of claim 26, wherein determining the cell type comprises determining a cell phenotype.

28. The method of claim 27, wherein determining the cell phenotype comprises determining a tissue type, a cancer type, a cancer subtype, a malignancy aggressiveness phenotype, and/or a drug responsivity phenotype.

29. The method of claim 27, wherein determining the cell phenotype includes determining expression of one or more genes of interest.

30. A method of detecting the presence of a cancer cell in a subject, comprising:

obtaining sequence read data generated from the sample comprising cell-free DNA obtained from the subject;

performing the computer-implemented method recited in any one of claims 5 to 21; and

determining the presence of a cancer cell in the subject based on the prediction provided by the computing system.

31. The method of claim 30, wherein the method is performed a plurality of times over time, wherein the detected cancer cell(s) in the subject at each performance of the method are further characterized to determine a cancer subtype or phenotype of the detected cancer cell(s) based on the prediction provided by the computing system.

32. The method of claim 31, wherein the method is performed a plurality of times over time, and wherein the method further comprises detecting a change in phenotype of the detected cancer cell(s) over time.

33. The method of claim 31 or 32, wherein the subject receives a cancer therapy between performances of the method, wherein the method further comprises determining the responsivity of the cancer cell(s) to the treatment.

34. A method of determining a cancer subtype of a target cancer cell from a sample comprising cell-free DNA derived from the target cancer cell, the method comprising:

obtaining sequence read data generated from the sample comprising cell-free DNA;

performing the computer-implemented method recited in any one of claims 5 to 21; and

determining the cell type of the originating cell based on the predicted cancer subtype provided by the computing system.

35. The method of claim 34, wherein the sample is obtained from a subject with cancer.

36. The method of any one of claims 30 to 35, wherein the cancer is characterized as metastatic breast cancer.

37. The method of claim 36, wherein determining the cancer subtype comprises determining whether the cancer is ER+ versus ER-.

38. The method of claim 36, wherein determining the cancer subtype comprises determining whether the cancer is PR+ versus PR-.

39. The method of claim 36, wherein determining the cancer subtype comprises determining whether the cancer is HER2+ versus HER2-.

40. The method of claim 36, wherein determining the cancer subtype comprises determining two or all of:

whether the cancer is ER+ versus ER-,

whether the cancer is PR+ versus PR-, and

whether the cancer is HER2+ versus HER2-.

41. The method of any one of claims 30 to 35, wherein the cancer is characterized as metastatic prostate cancer.

42. The method of claim 41, wherein determining the cancer subtype comprises determining whether the cancer is AR+ (ARPC) versus AR-.

43. The method of claim 41, wherein determining the cancer subtype comprises determining whether the cancer is ARPC versus AR-low.

44. The method of claim 41,wherein determining the cancer subtype comprises determining whether the cancer has a neuroendocrine prostate cancer (NEPC) phenotype signature or not.

45. The method of claim 41,wherein determining the cancer subtype comprises determining whether the cancer is amphicrine.

46. The method of claim 41, wherein determining the cancer subtype comprises determining two or all of:

whether the cancer is AR+ (ARPC) or AR-,

whether the cancer is AR-low or ARPC,

whether the cancer has a neuroendocrine prostate cancer (NEPC) phenotype signature or not,

whether the cancer is AR-low or NEPC,

whether the cancer is amphicrine or ARPC or NEPC.

47. The method of any one of claims 30 to 35, wherein the cancer is characterized as lung cancer.

48. The method of claim 47, wherein determining the cancer subtype comprises determining whether the cancer is small cell lung cancer (SCLC) or non-small cell lung cancer (NSCLC).

49. The method of claim 48, further comprising determining whether the NSCLC is adenocarcinoma or squamous cell carcinoma.

50. The method of any one of claims 47 to 49, wherein the sequence read data is generated from a panel of genomic targets.

51. The method of claim 50, wherein the panel of genomic targets comprises transcription factor binding sites (TFBSs) of one or more transcription factors associated with SCLC.

52. The method of claim 51, wherein the one or more transcription factors associated with SCLC comprise one or more of ASLC, NEUROD1, POU2F3, REST, and the like, and wherein the method comprises determining the nucleosome occupancy of the TFBSs.

53. The method of claim 52, wherein the TFBSs are identified by ChIP-seq data, or the like, and are retained in the panel if they are proximal to a transcription start site of a gene associated with lung cancer.

54. The method of claim 50, wherein the panel of genomic targets comprise transcription start sites (TSSs) for one or more markers associated with lung cancer, wherein the method comprises determining the nucleosome occupancy of the TSSs.

55. The method of any one of claims 30 to 54, wherein the sample is obtained from a subject and the method further comprises administering an effective treatment to the subject based on the determined cancer subtype.

56. The method of any one of claims 35 to 55, further comprising performing the method on a plurality of samples obtained from the subject at a plurality of distinct time points after an initial diagnosis of cancer.

57. The method of any one of claims 22 to 56, wherein the sequence read data is generated by ultra-low pass whole genome sequencing.

58. The method of any one of claims 22 to 56, wherein the sequence read data is generated by a chromatin accessibility assay.

59. The method of any one of claims 22 to 56, wherein the sequence read data is generated in an ATAC-seq method.

60. The method of any one of claims 22 to 56, wherein the sequence read data is generated in a ChIP-seq method.

61. The method of any one of claims 22 to 56, wherein the sequence read data is generated in a DNAse sensitivity assay.

62. The method of any one of claims 22 to 56, wherein the sequence read data is generated in a CUT&RUN assay.

63. The method of claim 62, wherein the CUT&RUN assay incorporates an affinity reagent that targets a post-translational modification to one or more of H3K27ac, H3K4me1 and H3K27ac.

64. The method of any one of claims 1 to 63, further comprising generating the sequence read data.

65. The method of any one of claims 1 to 64, wherein the sequence read data comprises sequence read data generated from a panel of genomic targets.

66. The method of claim 65, wherein the panel of genomic targets comprises transcription factor binding sites (TFBSs) of one or more transcription factors associated with a cancer type of interest.

67. The method of claim 66, wherein the method comprises determining the nucleosome occupancy of the TFBSs.

68. The method of claim 66, wherein the TFBSs are identified by ChIP-seq data, or the like, and are retained in the panel if they are proximal to a transcription start site of a gene associated with the cancer type of interest.

69. The method of claim 66, wherein the panel of genomic targets comprise transcription start sites (TSSs) for one or more markers associated with the cancer type of interest, wherein the method comprises determining the nucleosome occupancy of the TSSs.

70. The method of any one of claims 22 to 64, wherein the sample is blood, plasma, or serum.

71. A computer-implemented method of enhancing sequence read data from cell-free DNA samples for cell type prediction, the method comprising:

receiving, by a computing system, sequence read data, wherein the sequence read data includes a plurality of fragment reads, and wherein each fragment read has a fragment length;

determining, by the computing system, a fragment size variability for at least one gene associated with a cell type; and

predicting, by the computing system, the cell type based on the fragment size variability for the at least one gene.

72. The computer-implemented method of claim 71, wherein determining the fragment size variability includes determining a fragment size coefficient of variation.

73. The computer-implemented method of claim 71, wherein predicting the cell type based on the genomic coverage distribution includes predicting a cell phenotype.

74. The computer-implemented method of claim 73, wherein predicting the cell phenotype includes predicting a cancer subtype.

75. The computer-implemented method of claim 74, wherein predicting the cell phenotype includes predicting a cancer subtype of prostate cancer.

76. The computer-implemented method of claim 75, wherein predicting the cancer subtype includes distinguishing between ARPC and NEPC.

77. The computer-implemented method of claim 71, wherein predicting the cell type based on the fragment size variability includes:

generating one or more features based on the fragment size variability;

providing the one or more features as input to a classifier model; and

determining the cell type based on an output of the classifier model.

78. The computer-implemented method of claim 77, wherein generating the one or more features based on the fragment size variability includes generating a $\log_2$ fold change value of a fragment size coefficient of variation in a first cell type versus a second cell type.

79. The computer-implemented method of claim 78, wherein the $\log_2$ fold change value predicts at least one of gene expression and gene transcriptional activity between the first cell type and the second cell type.

80. The computer-implemented method of claim 78, wherein the first cell type is an ARPC cell and the second cell type is an NEPC cell.

81. The computer-implemented method of claim 77, wherein the classifier model includes a logistic regression model, an artificial neural network, a decision tree, a support vector machine, or a Bayesian network.

82. A method for determining a cell type of a cell of interest from a sample comprising cell-free DNA derived from the cell of interest, comprising:

obtaining sequence read data generated from the sample comprising cell-free DNA;

performing the computer-implemented method recited in any one of claims 71 to 81; and

determining the cell type of the cell of interest based on the prediction provided by the computing system.

83. The method of claim 82, wherein determining the cell type comprises determining a cell phenotype.

84. The method of claim 83, wherein determining the cell phenotype comprises determining a cancer subtype.

85. The method of claim 84, wherein determining the cancer subtype includes distinguishing between ARPC and NEPC.

86. A method of detecting the presence of a cancer cell in a subject, comprising:

obtaining sequence read data generated from a sample comprising cell-free DNA obtained from the subject;

performing the computer-implemented method recited in any one of claims 71 to 81; and

determining the presence of a cancer cell in the subject based on the prediction provided by the computing system.

87. The method of claim 86, wherein the method is performed a plurality of times over time, wherein the detected cancer cell(s) in the subject at each performance of the method are further characterized to determine a cancer subtype or phenotype of the detected cancer cell(s) based on the prediction provided by the computing system.

88. The method of claim 87, wherein the method is performed a plurality of times over time, and wherein the method further comprises detecting a change in phenotype of the detected cancer cell(s) over time.

89. The method of claim 87 or 88, wherein the subject receives a cancer therapy between performances of the method, wherein the method further comprises determining the responsivity of the cancer cell(s) to the treatment.

90. A method of determining a cancer subtype of a target cancer cell from a sample comprising cell-free DNA derived from the target cancer cell, the method comprising:

obtaining sequence read data generated from the sample comprising cell-free DNA;

performing the computer-implemented method recited in any one of claims 71 to 81; and

determining the cell type of the originating cell based on the predicted cancer subtype provided by the computing system.

91. The method of claim 90, wherein the sample is obtained from a subject with cancer.

92. The method of any one of claims 86 to 91, wherein the cancer is characterized as metastatic prostate cancer.

93. The method of claim 92, wherein the determining the cancer subtype comprises determining whether the cancer is AR+ (ARPC) versus AR-.

94. The method of claim 92, wherein the determining the cancer subtype comprises determining whether the cancer is ARPC versus AR-low prostate cancer (ARLPC).

95. The method of claim 92, wherein the determining the cancer subtype comprises determining whether the cancer has a neuroendocrine prostate cancer (NEPC) phenotype signature or not.

96. The method of any one of claims 86 to 95, wherein the sample is obtained from a subject and the method further comprises administering an effective treatment to the subject based on the determined cancer subtype.

97. The method of any one of claims 86 to 96, further comprising performing the method on a plurality of samples obtained from the subject at a plurality of distinct time points after an initial diagnosis of cancer.

98. The method of any one of any one of claims 82 to 97, wherein the sequence read data is generated by ultra-low pass whole genome sequencing.

99. The method of any one of claims 82 to 97, wherein the sequence read data is generated by a chromatin accessibility assay.

100. The method of any one of claims 82 to 97, wherein the sequence read data is generated in an ATAC-seq method.

101. The method of any one of claims 82 to 97, wherein the sequence read data is generated in a ChIP-seq method.

102. The method of any one of claims 82 to 97, wherein the sequence read data is generated in a DNAse sensitivity assay.

103. The method of any one of claims 82 to 97, wherein the sequence read data is generated in a CUT&RUN assay.

104. The method of claim 103, wherein the CUT&RUN assay incorporates an affinity reagent that targets a post-translational modification to one or more of H3K27ac, H3K4me1 and H3K27ac.

105. The method of any one of claims 71 to 104, further comprising generating the sequence read data.

106. The method of any one of claims 71 to 105, wherein the sequence read data is generated from a panel of genomic targets.

107. The method of claim 106, wherein the panel of genomic targets comprises transcription factor binding sites (TFBSs) of one or more transcription factors associated with a cancer type of interest.

108. The method of claim 107, wherein the method comprises determining the nucleosome occupancy of the TFBSs.

109. The method of claim 107, wherein the TFBSs are identified by ChIP-seq data, or the like, and are retained in the panel if they are proximal to a transcription start site of a gene associated with the cancer type of interest.

110. The method of claim 107, wherein the panel of genomic targets comprise transcription start sites (TSSs) for one or more markers associated with the cancer type of interest, wherein the method comprises determining the nucleosome occupancy of the TSSs.

111. The method of any one of claims 82 to 110, wherein the sample is blood, plasma, or serum.

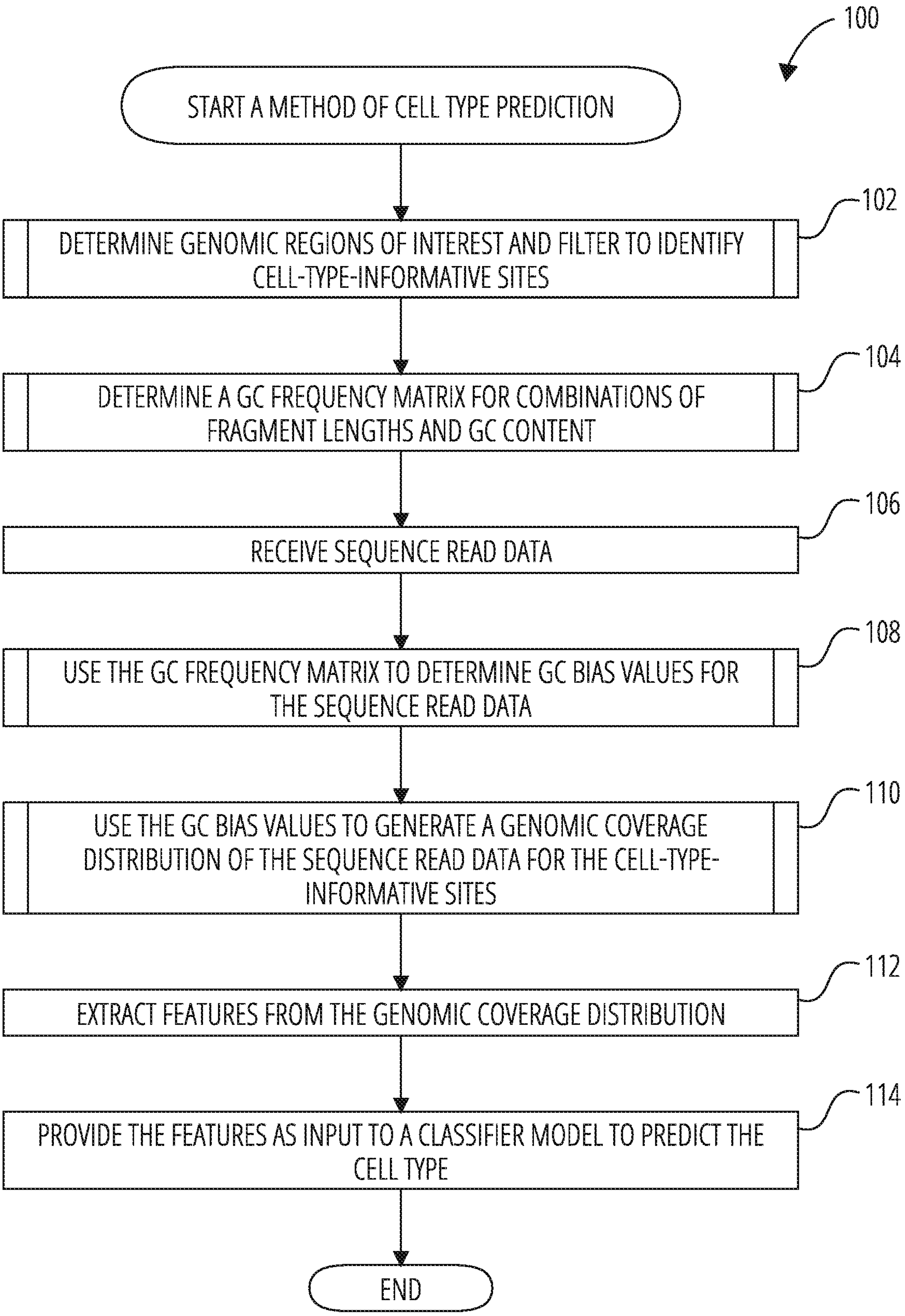

100
START A METHOD OF CELL TYPE PREDICTION
102
DETERMINE GENOMIC REGIONS OF INTEREST AND FILTER TO IDENTIFY CELL-TYPE-INFORMATIVE SITES
104
DETERMINE A GC FREQUENCY MATRIX FOR COMBINATIONS OF FRAGMENT LENGTHS AND GC CONTENT
106
RECEIVE SEQUENCE READ DATA
108
USE THE GC FREQUENCY MATRIX TO DETERMINE GC BIAS VALUES FOR THE SEQUENCE READ DATA
110
USE THE GC BIAS VALUES TO GENERATE A GENOMIC COVERAGE DISTRIBUTION OF THE SEQUENCE READ DATA FOR THE CELL-TYPE-INFORMATIVE SITES
112
EXTRACT FEATURES FROM THE GENOMIC COVERAGE DISTRIBUTION
114
PROVIDE THE FEATURES AS INPUT TO A CLASSIFIER MODEL TO PREDICT THE CELL TYPE
END

*FIG. 1*

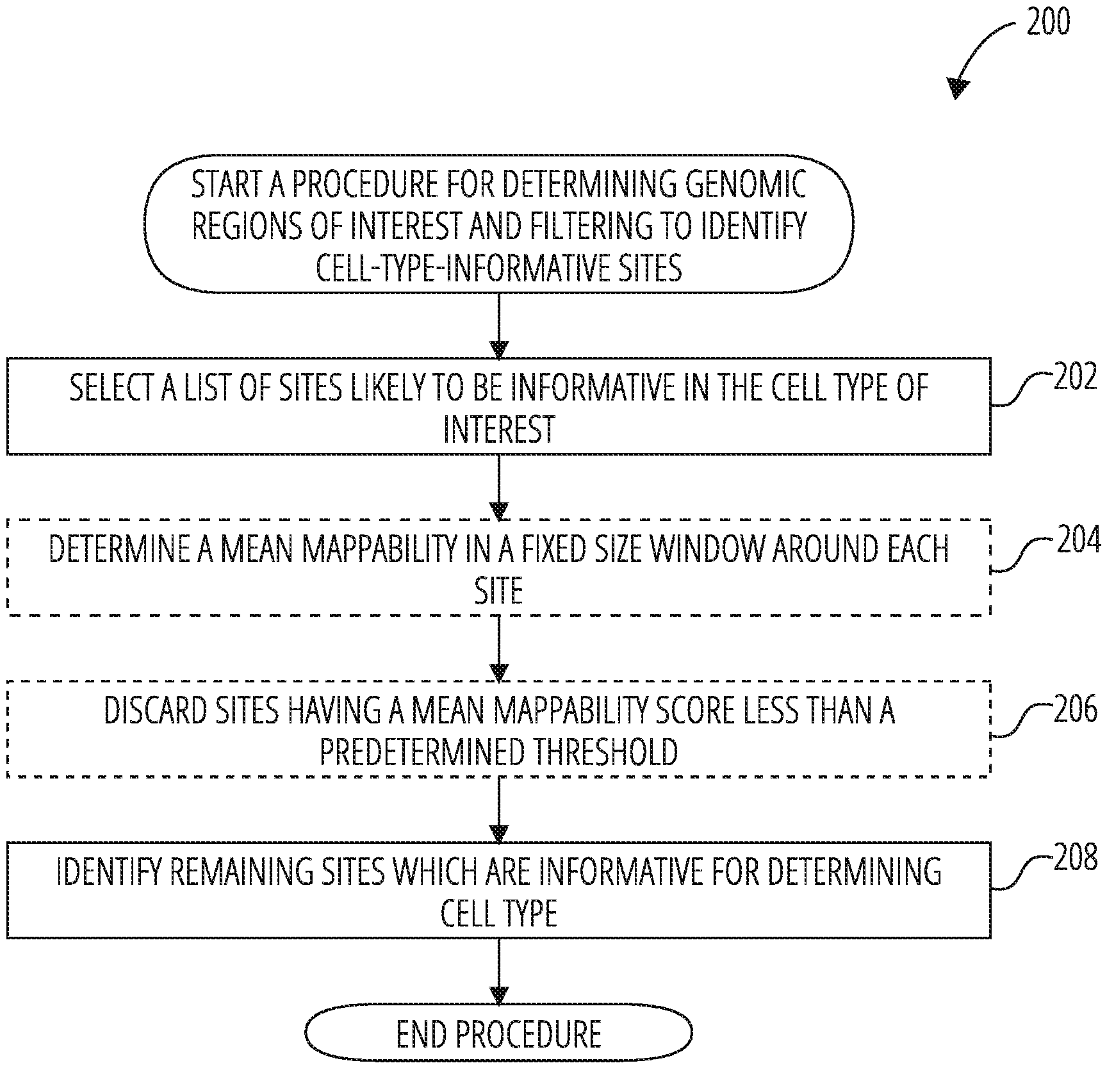

200
START A PROCEDURE FOR DETERMINING GENOMIC REGIONS OF INTEREST AND FILTERING TO IDENTIFY CELL-TYPE-INFORMATIVE SITES
SELECT A LIST OF SITES LIKELY TO BE INFORMATIVE IN THE CELL TYPE OF INTEREST
202
DETERMINE A MEAN MAPPABILITY IN A FIXED SIZE WINDOW AROUND EACH SITE
204
DISCARD SITES HAVING A MEAN MAPPABILITY SCORE LESS THAN A PREDETERMINED THRESHOLD
206
IDENTIFY REMAINING SITES WHICH ARE INFORMATIVE FOR DETERMINING CELL TYPE
208
END PROCEDURE

*FIG. 2*

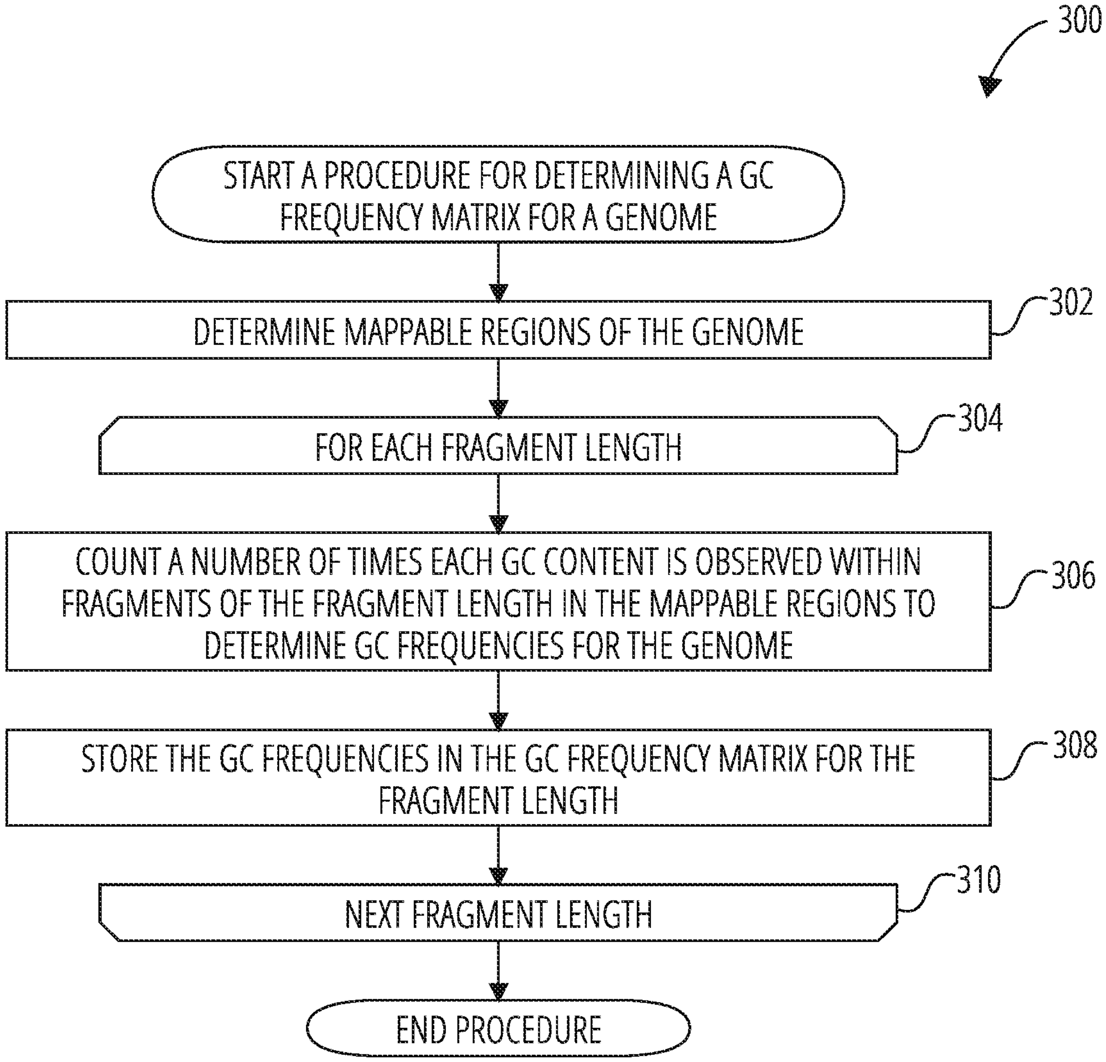

300
START A PROCEDURE FOR DETERMINING A GC FREQUENCY MATRIX FOR A GENOME
DETERMINE MAPPABLE REGIONS OF THE GENOME
302
FOR EACH FRAGMENT LENGTH
304
COUNT A NUMBER OF TIMES EACH GC CONTENT IS OBSERVED WITHIN FRAGMENTS OF THE FRAGMENT LENGTH IN THE MAPPABLE REGIONS TO DETERMINE GC FREQUENCIES FOR THE GENOME
306
STORE THE GC FREQUENCIES IN THE GC FREQUENCY MATRIX FOR THE FRAGMENT LENGTH
308
NEXT FRAGMENT LENGTH
310
END PROCEDURE

*FIG. 3*

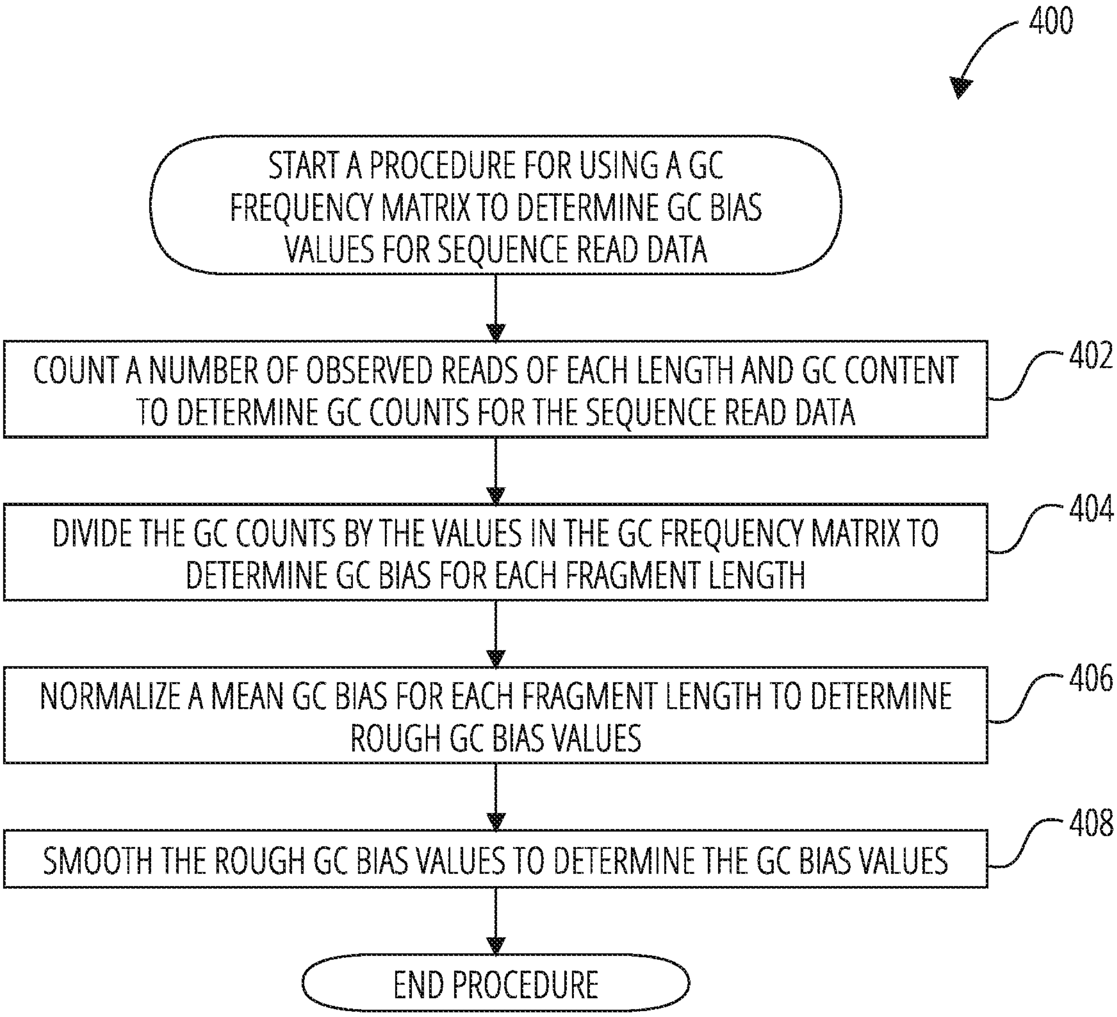

400
START A PROCEDURE FOR USING A GC FREQUENCY MATRIX TO DETERMINE GC BIAS VALUES FOR SEQUENCE READ DATA
COUNT A NUMBER OF OBSERVED READS OF EACH LENGTH AND GC CONTENT TO DETERMINE GC COUNTS FOR THE SEQUENCE READ DATA
402
DIVIDE THE GC COUNTS BY THE VALUES IN THE GC FREQUENCY MATRIX TO DETERMINE GC BIAS FOR EACH FRAGMENT LENGTH
404
NORMALIZE A MEAN GC BIAS FOR EACH FRAGMENT LENGTH TO DETERMINE ROUGH GC BIAS VALUES
406
SMOOTH THE ROUGH GC BIAS VALUES TO DETERMINE THE GC BIAS VALUES
408
END PROCEDURE

FIG. 4

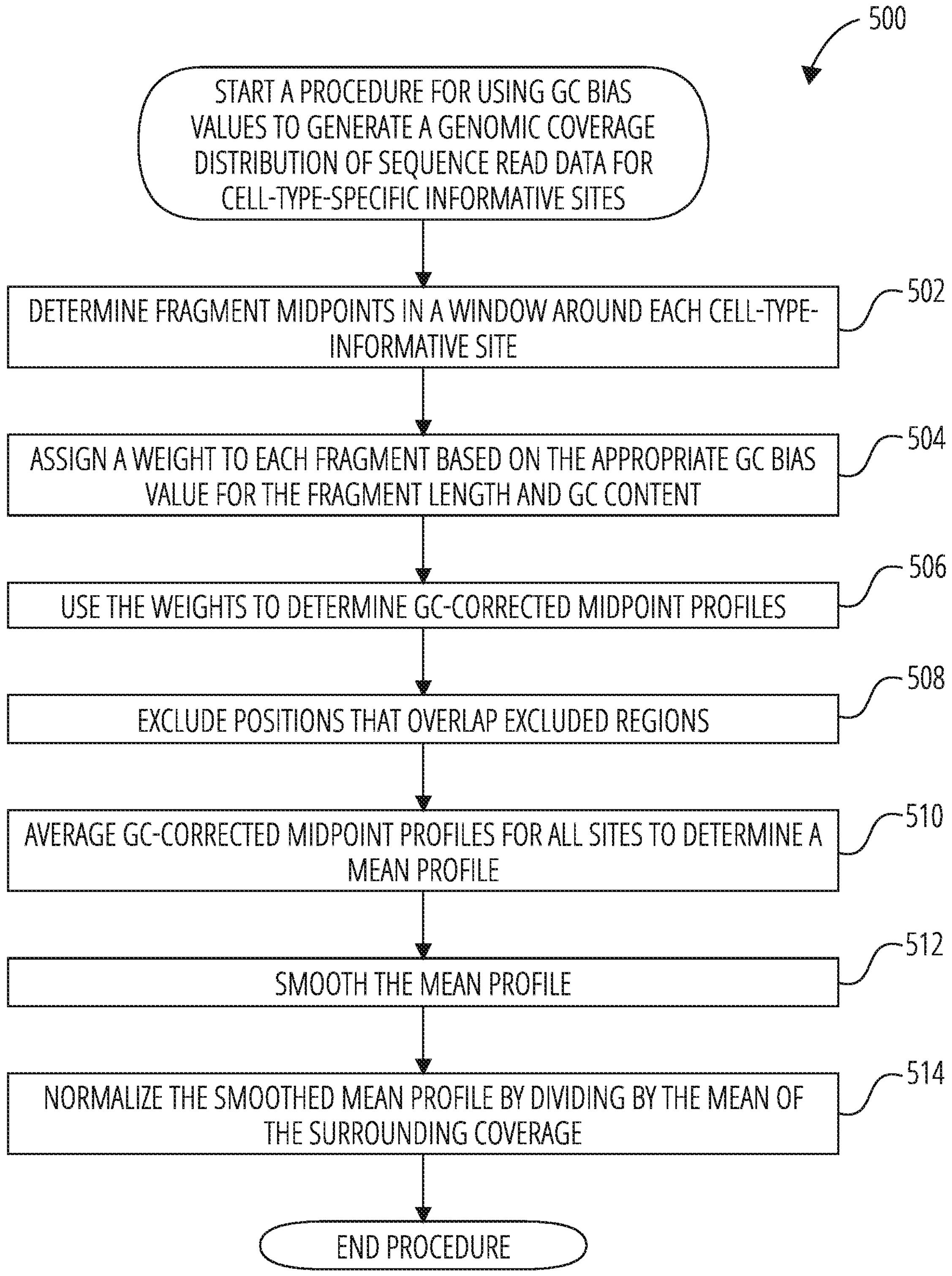

500
START A PROCEDURE FOR USING GC BIAS VALUES TO GENERATE A GENOMIC COVERAGE DISTRIBUTION OF SEQUENCE READ DATA FOR CELL-TYPE-SPECIFIC INFORMATIVE SITES
DETERMINE FRAGMENT MIDPOINTS IN A WINDOW AROUND EACH CELL-TYPE-INFORMATIVE SITE
502
ASSIGN A WEIGHT TO EACH FRAGMENT BASED ON THE APPROPRIATE GC BIAS VALUE FOR THE FRAGMENT LENGTH AND GC CONTENT
504
USE THE WEIGHTS TO DETERMINE GC-CORRECTED MIDPOINT PROFILES
506
EXCLUDE POSITIONS THAT OVERLAP EXCLUDED REGIONS
508
AVERAGE GC-CORRECTED MIDPOINT PROFILES FOR ALL SITES TO DETERMINE A MEAN PROFILE
510
SMOOTH THE MEAN PROFILE
512
NORMALIZE THE SMOOTHED MEAN PROFILE BY DIVIDING BY THE MEAN OF THE SURROUNDING COVERAGE
514
END PROCEDURE

*FIG. 5*

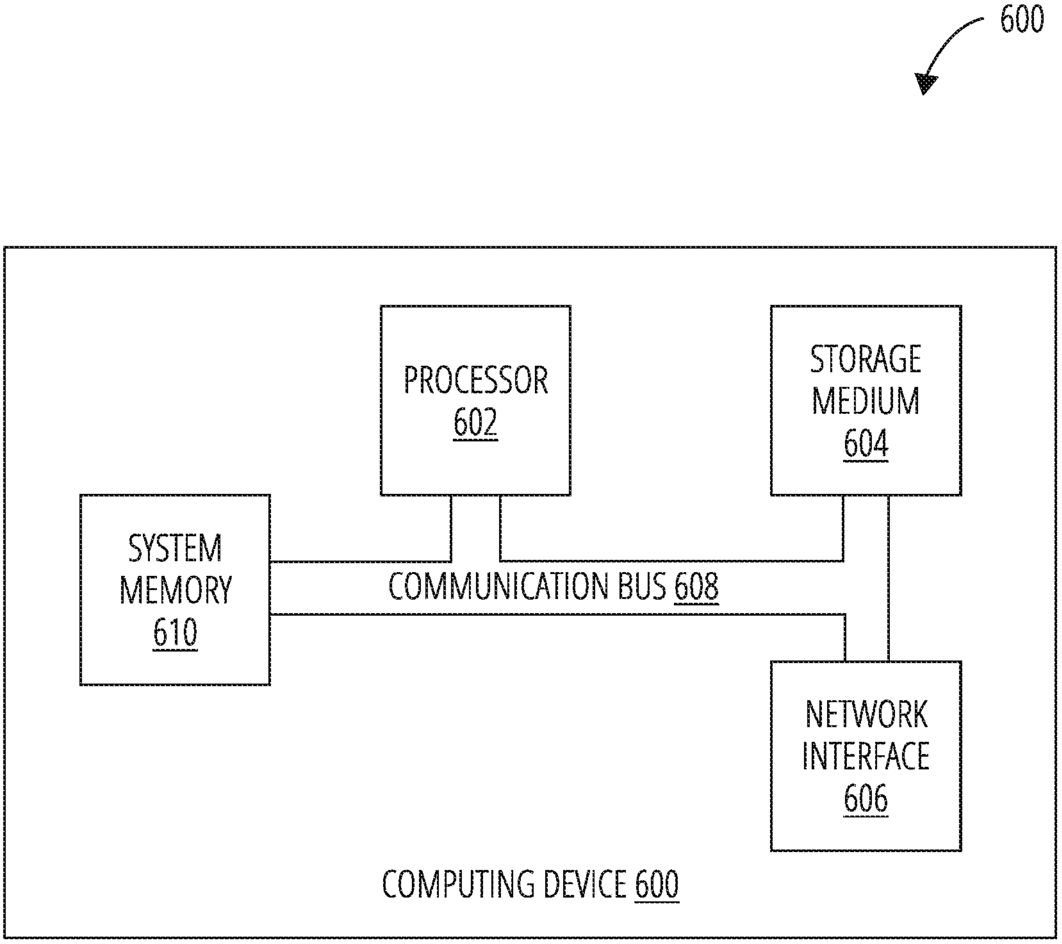

600
PROCESSOR 602
STORAGE MEDIUM 604
SYSTEM MEMORY 610
COMMUNICATION BUS 608
NETWORK INTERFACE 606
COMPUTING DEVICE 600

FIG. 6

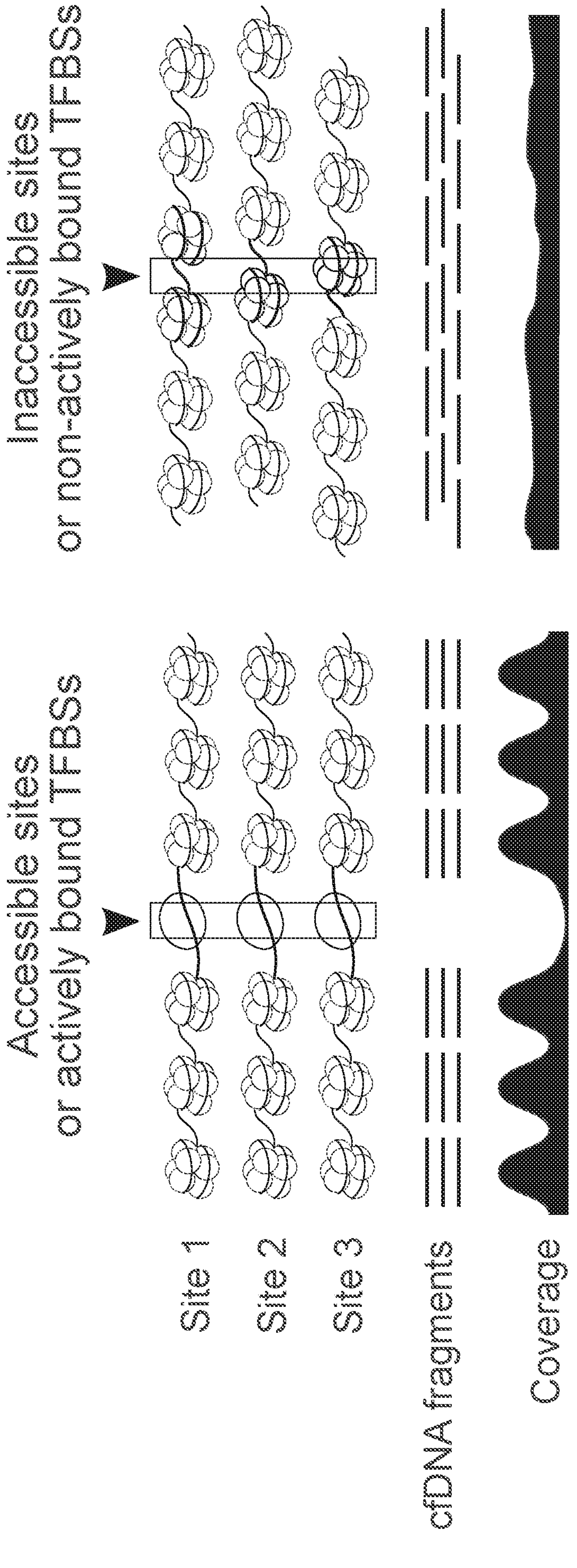

Accessible sites
or actively bound TFBSs
Inaccessible sites
or non-actively bound TFBSs
Site 1
Site 2
Site 3
cfDNA fragments
Coverage

FIG. 7A

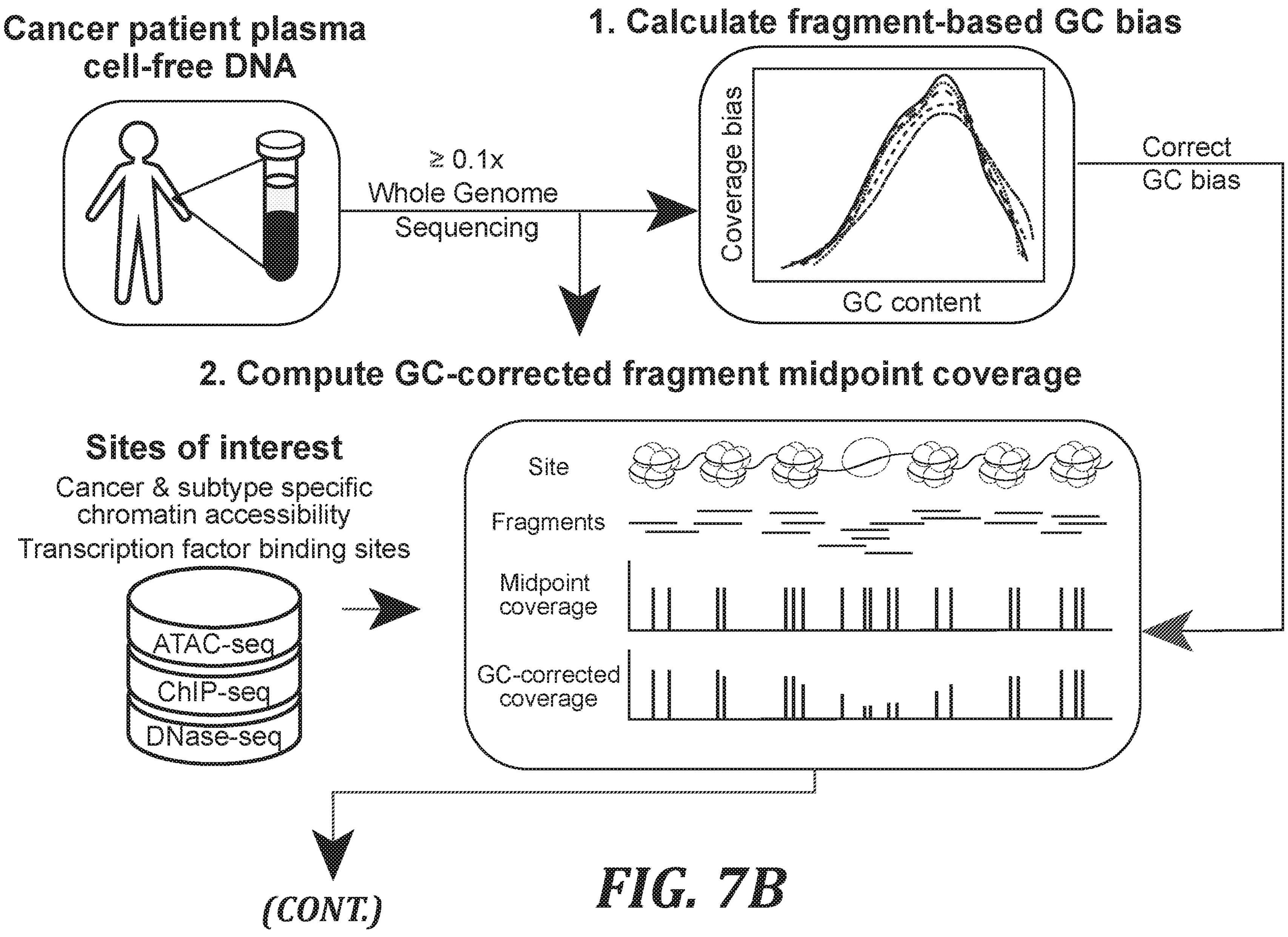

Cancer patient plasma cell-free DNA
≥ 0.1x
Whole Genome
Sequencing
1. Calculate fragment-based GC bias
Coverage bias
GC content
Correct
GC bias
2. Compute GC-corrected fragment midpoint coverage
Sites of interest
Cancer & subtype specific
chromatin accessibility
Transcription factor binding sites
ATAC-seq
ChIP-seq
DNase-seq
Site
Fragments
Midpoint
coverage
GC-corrected
coverage
(CONT.)

FIG. 7B

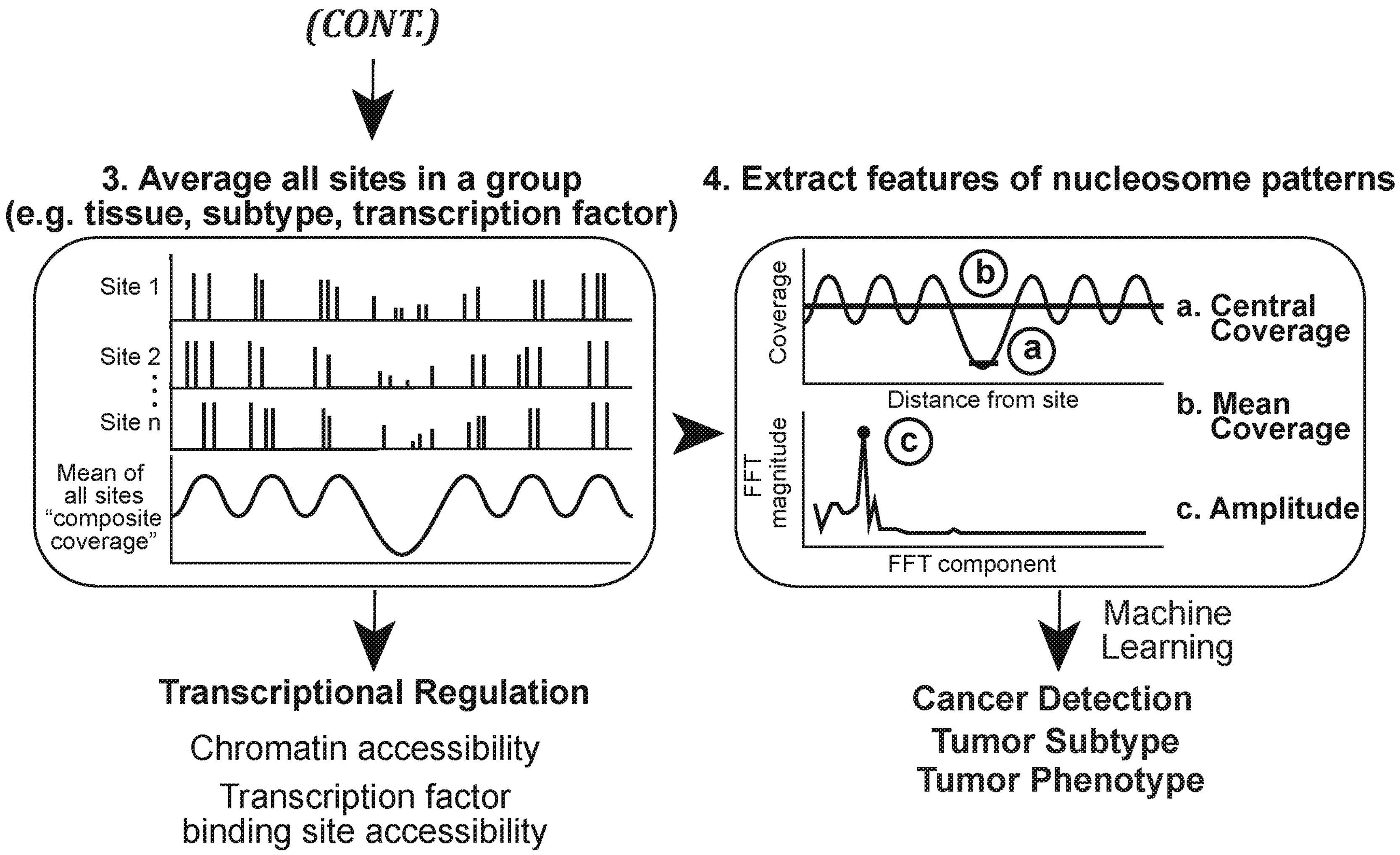

(CONT.)
3. Average all sites in a group
(e.g. tissue, subtype, transcription factor)
Site 1
Site 2
Site n
Mean of all sites "composite coverage"
4. Extract features of nucleosome patterns
Coverage
b
a
Distance from site
FFT magnitude
c
FFT component
a. Central Coverage
b. Mean Coverage
c. Amplitude
Transcriptional Regulation
Chromatin accessibility
Transcription factor binding site accessibility
Machine Learning
Cancer Detection
Tumor Subtype
Tumor Phenotype

FIG. 7B
(CONT.)

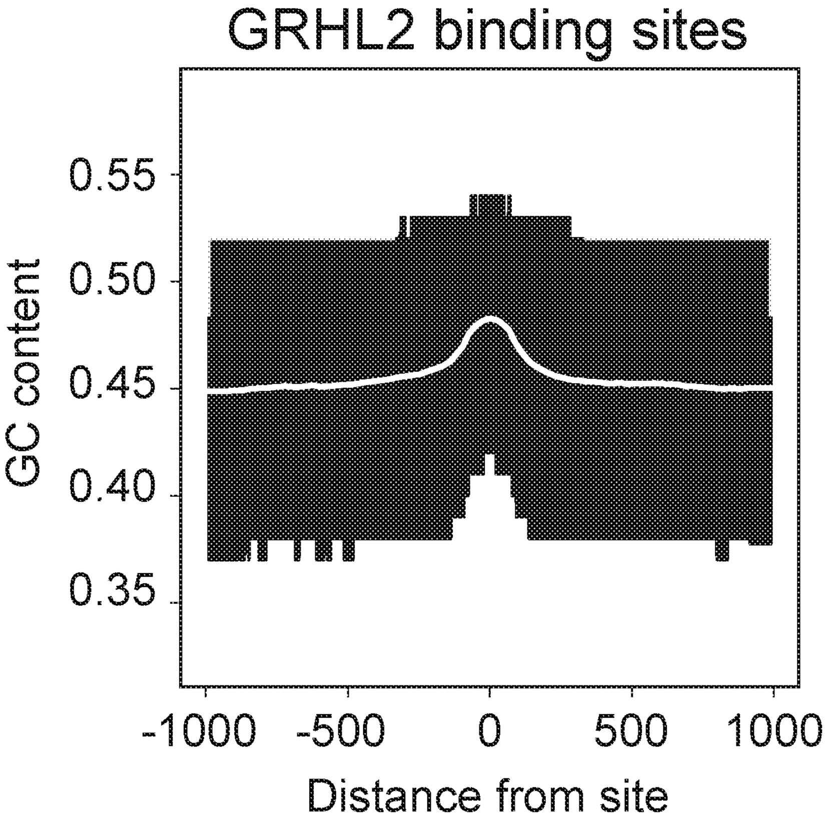

GRHL2 binding sites
GC content
0.55
0.50
0.45
0.40
0.35
-1000
-500
0
500
1000
Distance from site

*FIG. 8A*

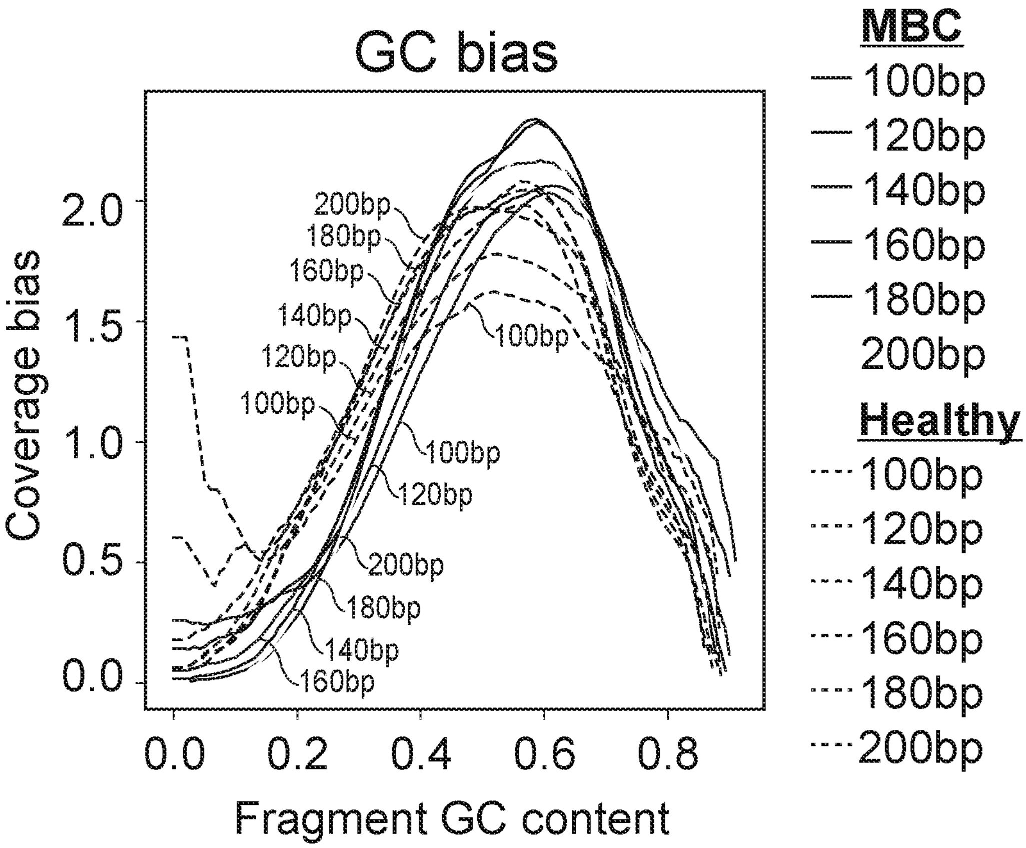

GC bias
Coverage bias
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
Fragment GC content
200bp
180bp
160bp
140bp
120bp
100bp
100bp
100bp
120bp
200bp
180bp
140bp
160bp
MBC
100bp
120bp
140bp
160bp
180bp
200bp
Healthy
100bp
120bp
140bp
160bp
180bp
200bp

*FIG. 8B*

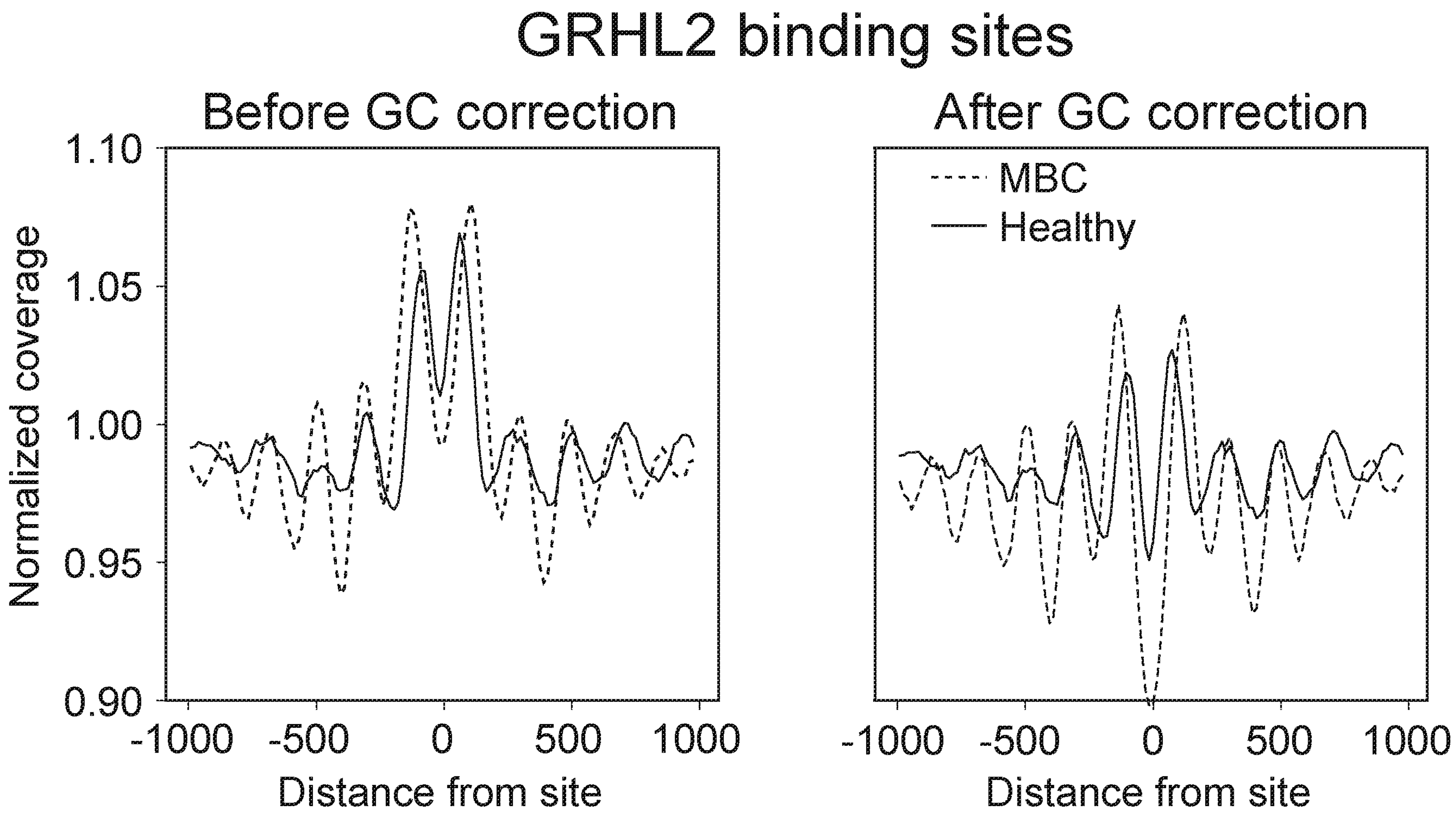

GRHL2 binding sites
Before GC correction
After GC correction
Normalized coverage
1.10
1.05
1.00
0.95
0.90
-1000
-500
0
500
1000
Distance from site
MBC
Healthy

*FIG. 8C*

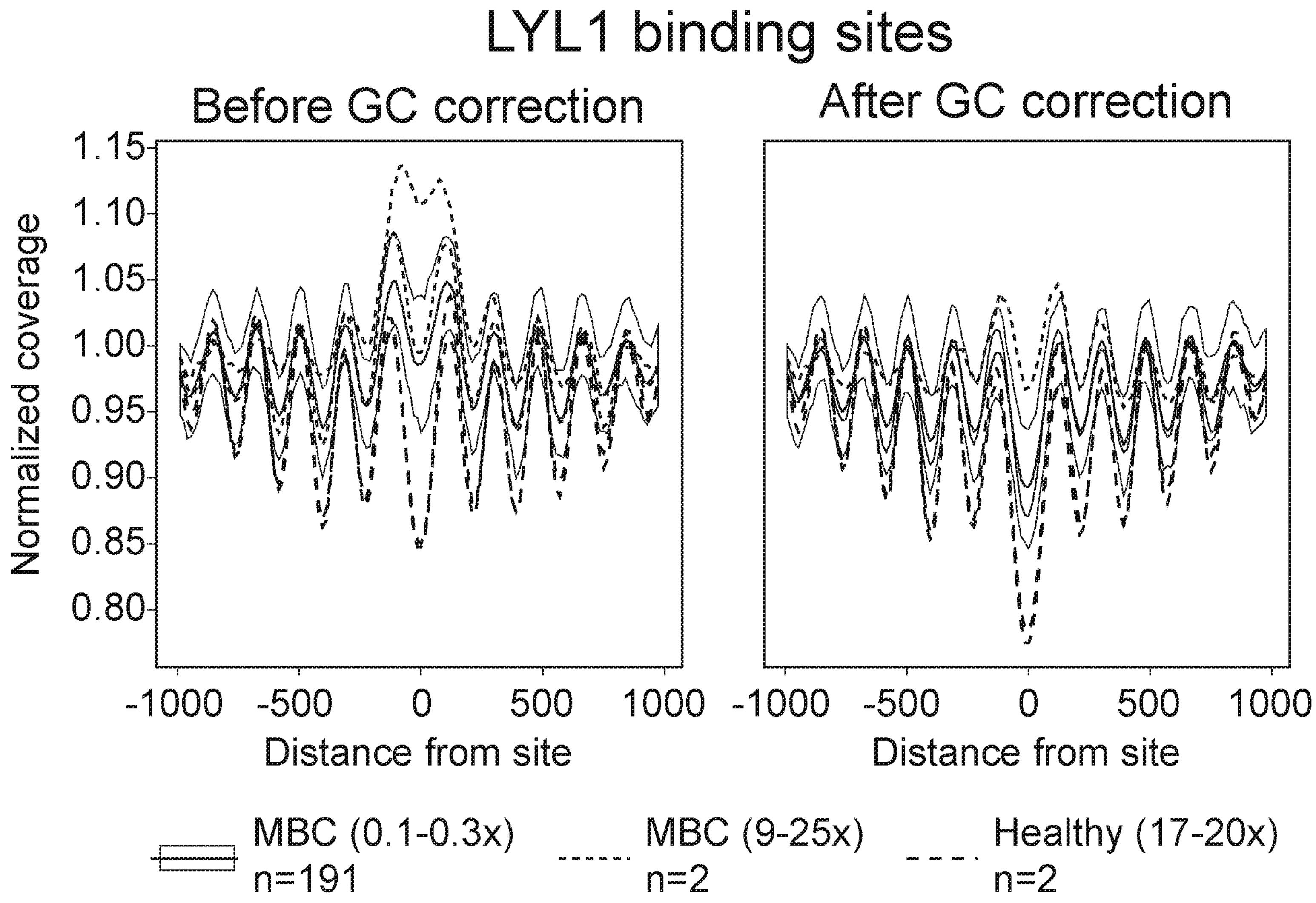

LYL1 binding sites
Before GC correction
After GC correction
Normalized coverage
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.80
-1000
-500
0
500
1000
Distance from site
MBC (0.1-0.3x)
n=191
MBC (9-25x)
n=2
Healthy (17-20x)
n=2

FIG. 8D

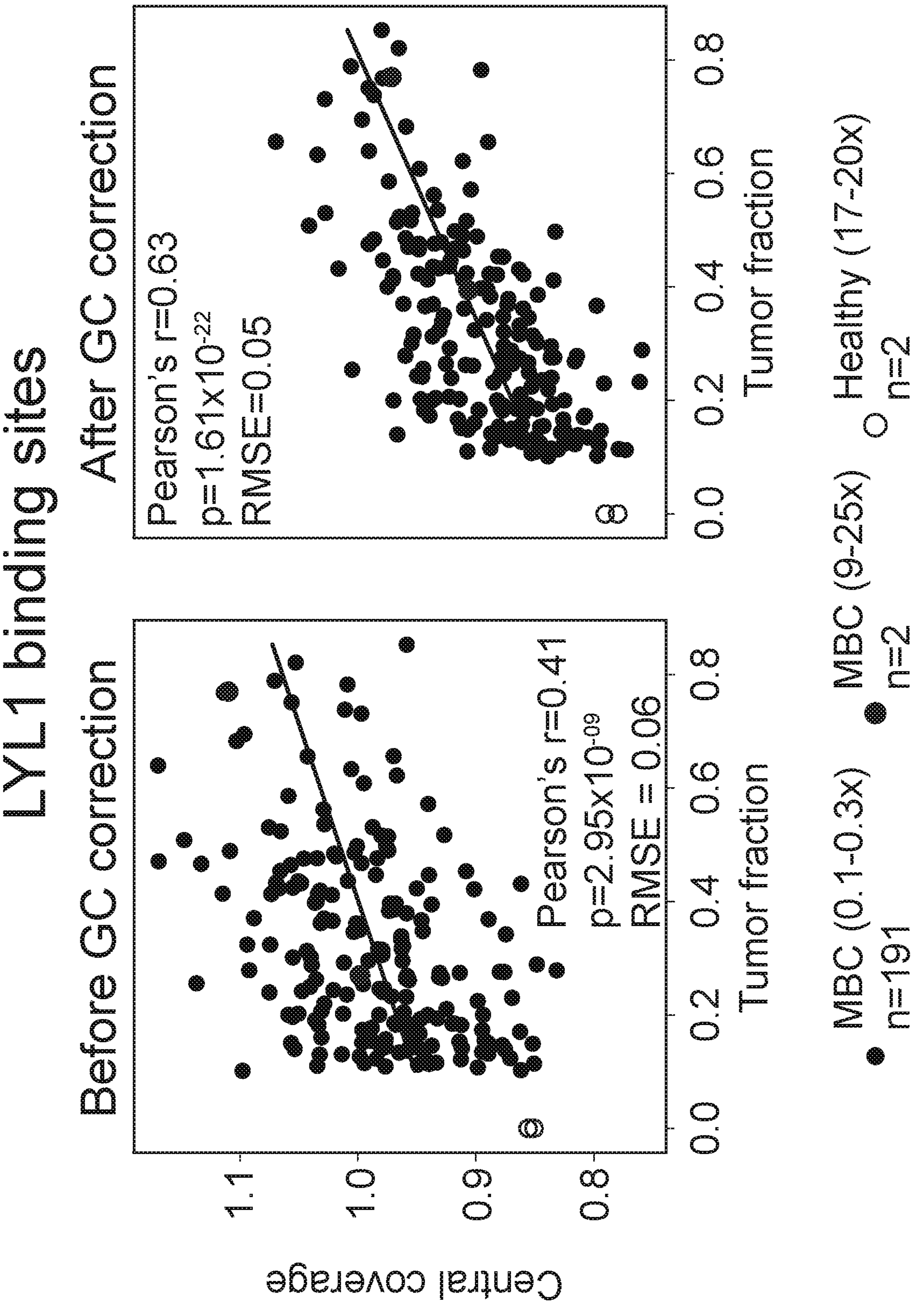

LYL1 binding sites
Before GC correction
After GC correction
Central coverage
1.1
1.0
0.9
0.8
0.0
0.2
0.4
0.6
0.8
Tumor fraction
Pearson's r=0.41
p=2.95x10-09
RMSE = 0.06
Pearson's r=0.63
p=1.61x10-22
RMSE=0.05
● MBC (0.1-0.3x) n=191
● MBC (9-25x) n=2
○ Healthy (17-20x) n=2

FIG. 8E

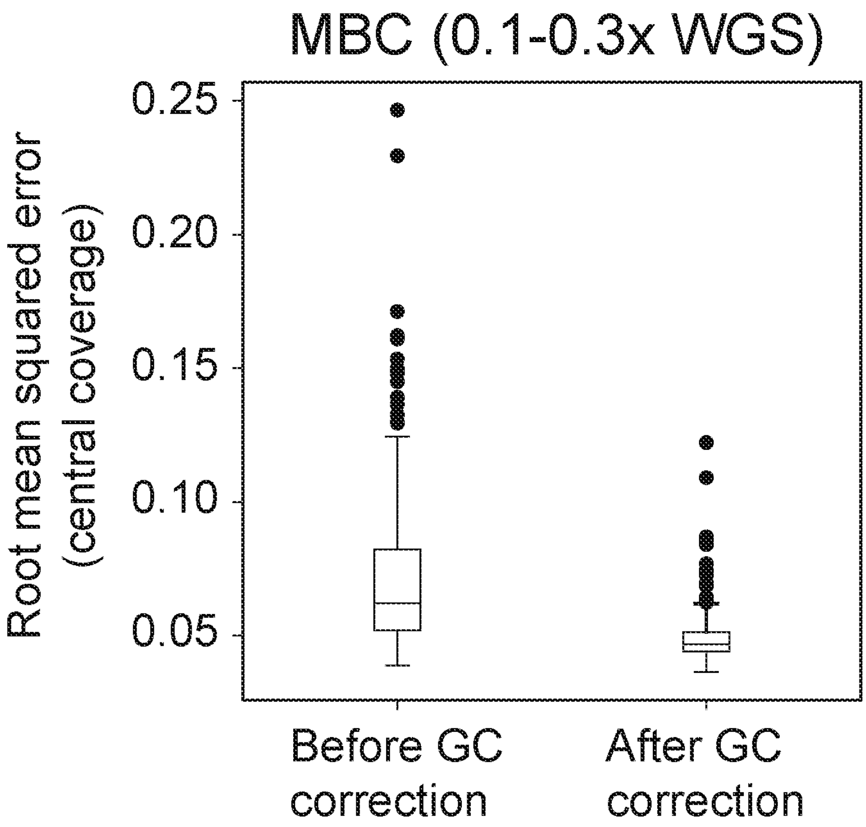

MBC (0.1-0.3x WGS)
Root mean squared error (central coverage)
0.25
0.20
0.15
0.10
0.05
Before GC correction
After GC correction

FIG. 8F

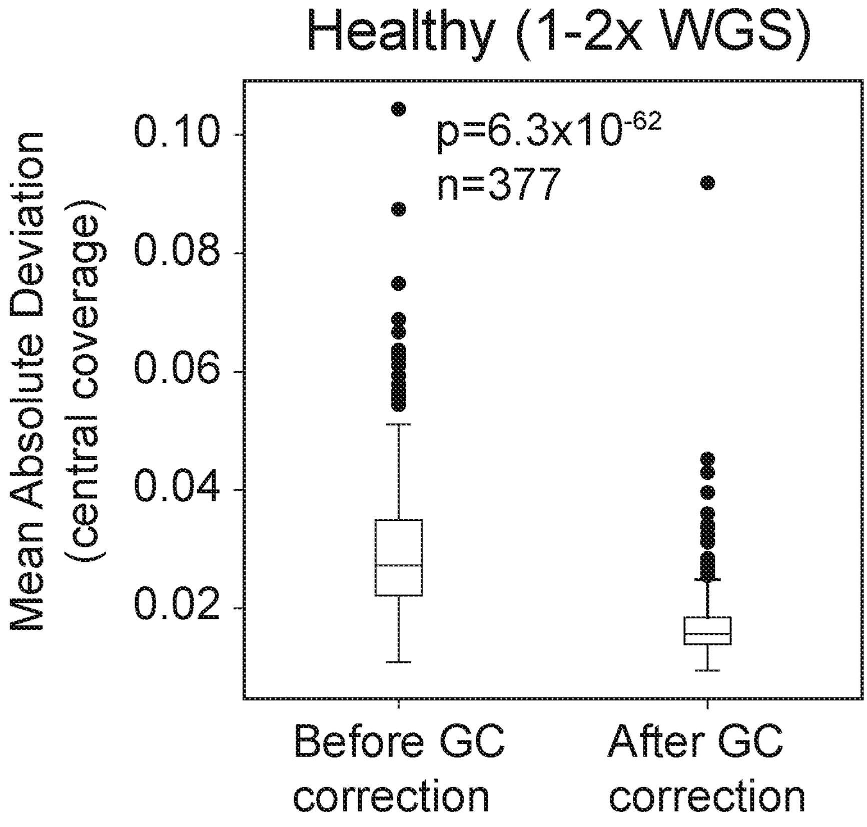

Healthy (1-2x WGS)
Mean Absolute Deviation (central coverage)
0.10
0.08
0.06
0.04
0.02
p=6.3x10^-62
n=377
Before GC correction
After GC correction

FIG. 8G

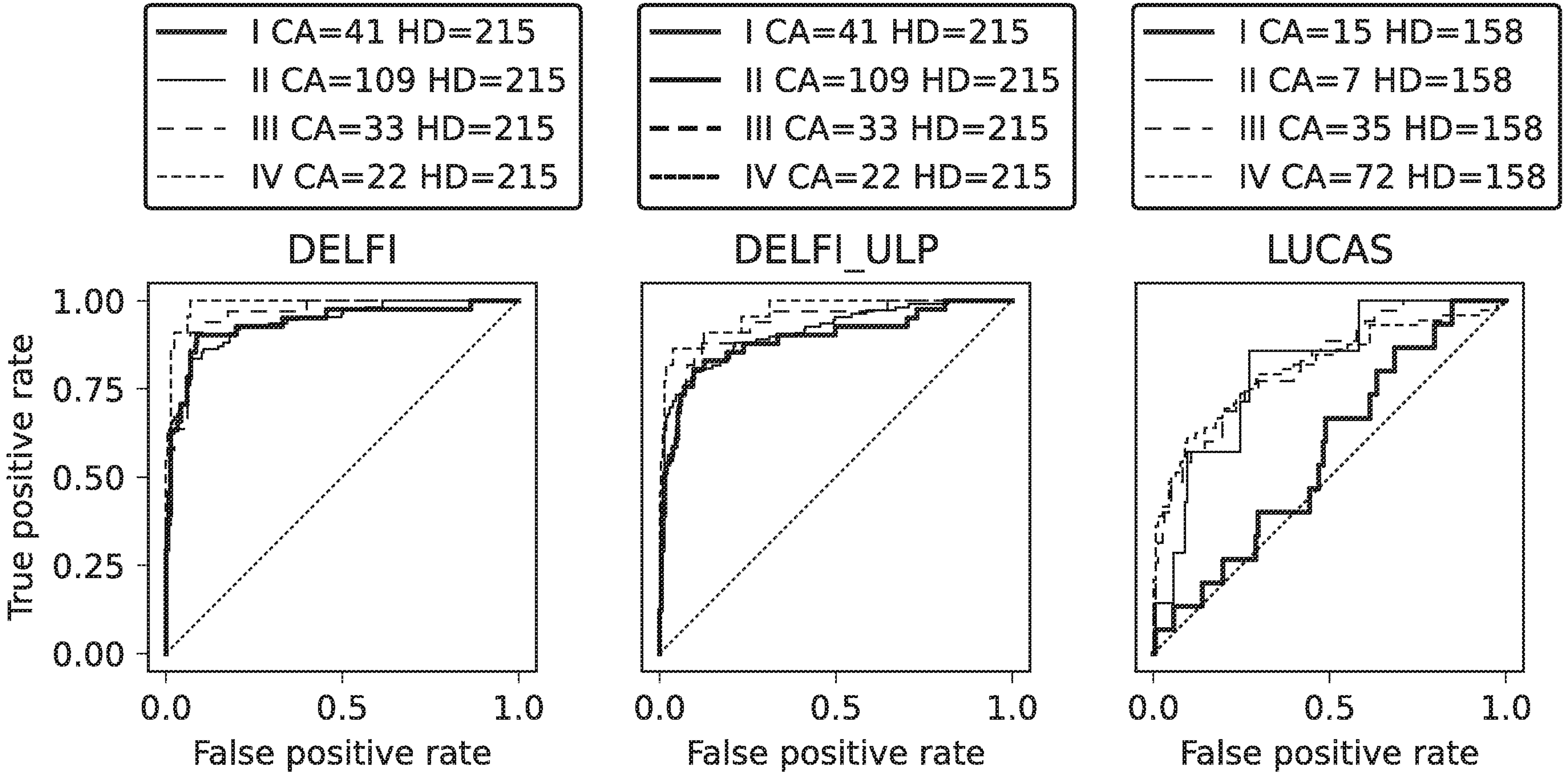

I CA=41 HD=215
II CA=109 HD=215
III CA=33 HD=215
IV CA=22 HD=215
DELFI
I CA=41 HD=215
II CA=109 HD=215
III CA=33 HD=215
IV CA=22 HD=215
DELFI_ULP
I CA=15 HD=158
II CA=7 HD=158
III CA=35 HD=158
IV CA=72 HD=158
LUCAS
True positive rate
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
False positive rate

FIG. 9A

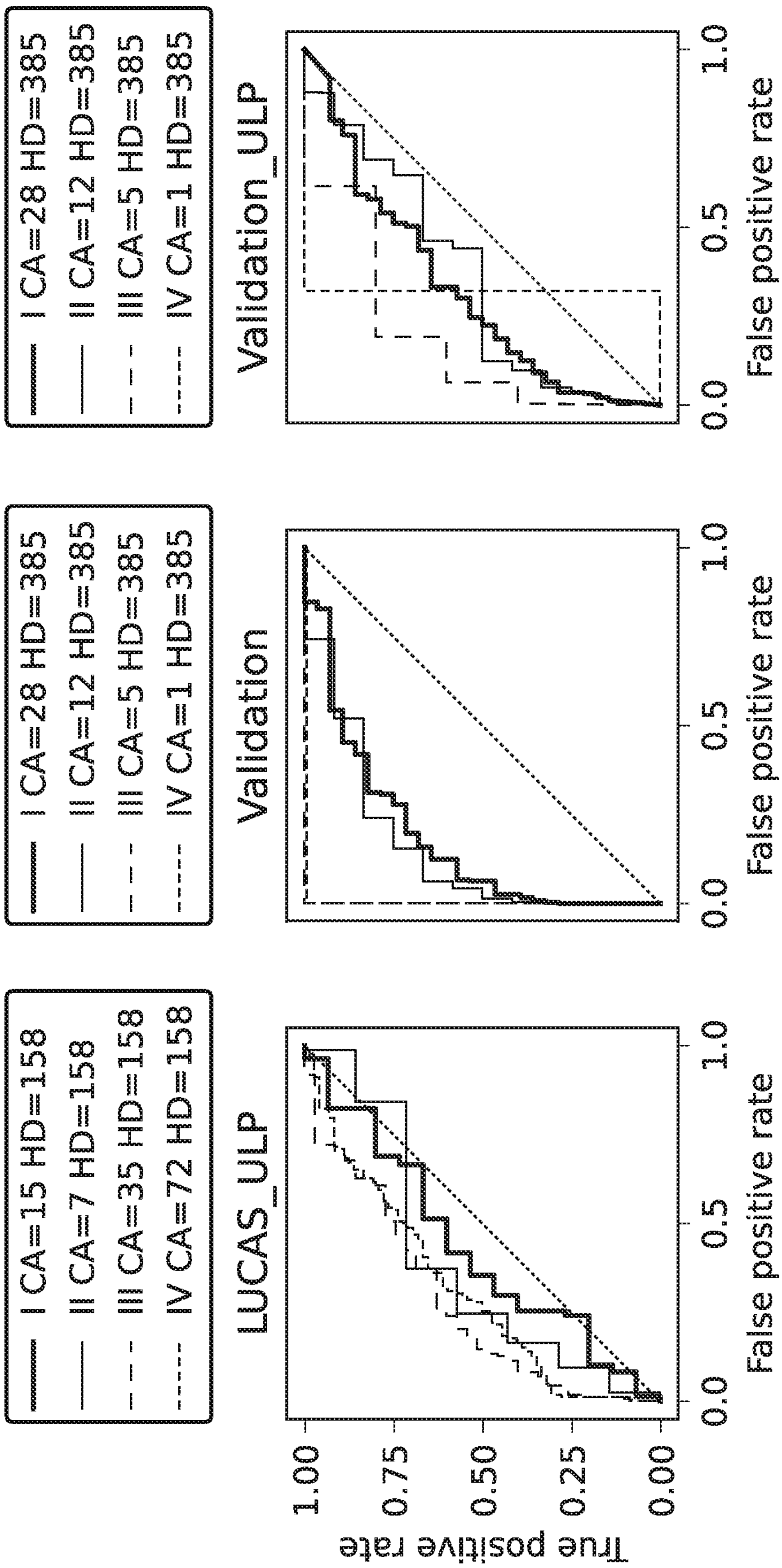

LUCAS_ULP
I CA=15 HD=158
II CA=7 HD=158
III CA=35 HD=158
IV CA=72 HD=158
True positive rate
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
False positive rate
Validation
I CA=28 HD=385
II CA=12 HD=385
III CA=5 HD=385
IV CA=1 HD=385
0.0
0.5
1.0
False positive rate
Validation_ULP
I CA=28 HD=385
II CA=12 HD=385
III CA=5 HD=385
IV CA=1 HD=385
0.0
0.5
1.0
False positive rate

FIG. 9A
(CONT.)

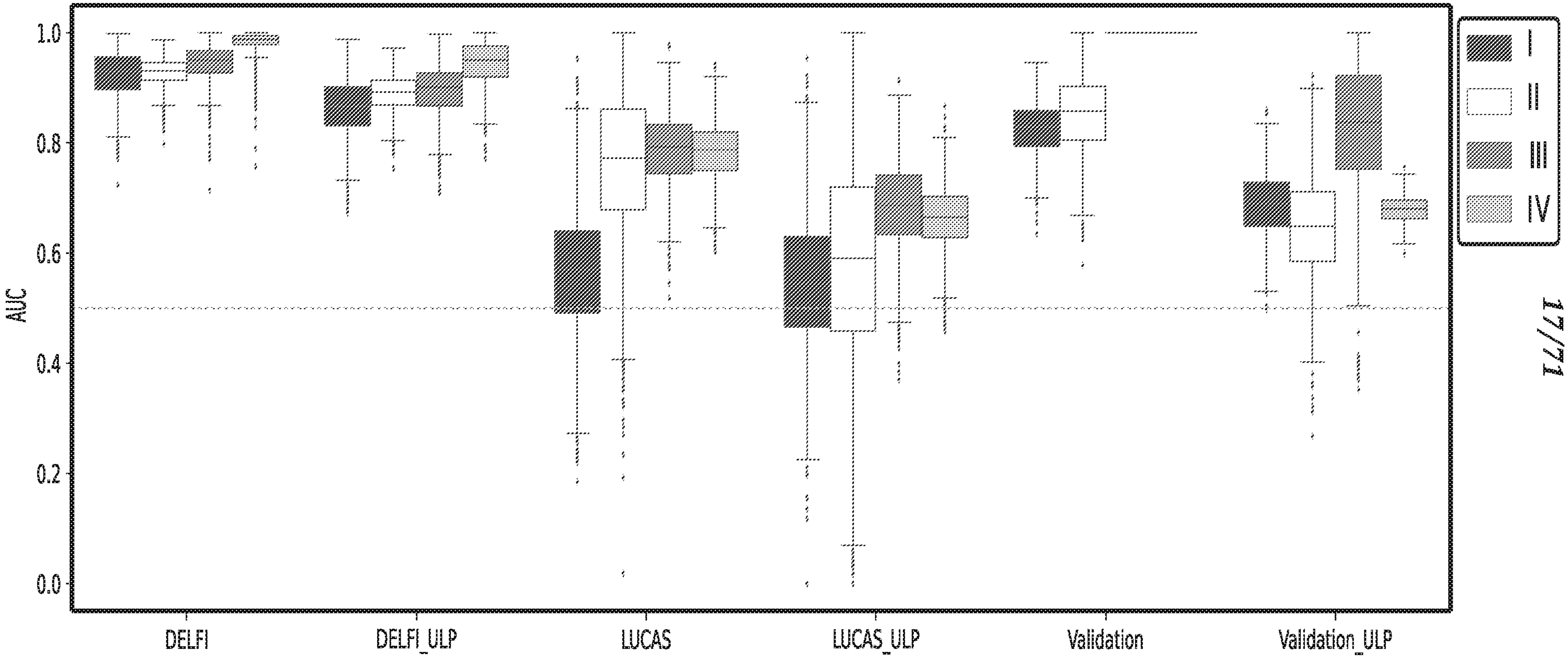

1.0
0.8
0.6
0.4
0.2
0.0
AUC
DELFI
DELFI_ULP
LUCAS
LUCAS_ULP
Validation
Validation_ULP
▨ I
□ II
▨ III
▨ IV

FIG. 9B

| | DELFI | DELFI_ULP | LUCAS | LUCAS_ULP | Validation | Validation_ULP |
|---|---|---|---|---|---|---|
| I | 0.93 (0.83-0.98) | 0.87 (0.75-0.95) | 0.57 (0.31-0.79) | 0.55 (0.28-0.79) | 0.83 (0.73-0.91) | 0.69 (0.57-0.8) |
| II | 0.93 (0.87-0.97) | 0.89 (0.82-0.95) | 0.77 (0.4-0.98) | 0.59 (0.09-0.95) | 0.86 (0.7-0.97) | 0.65 (0.45-0.83) |
| III | 0.95 (0.86-0.99) | 0.9 (0.78-0.97) | 0.79 (0.66-0.9) | 0.69 (0.51-0.83) | 1.0 (1.0-1.0) | 0.84 (0.57-1.0) |
| IV | 0.99 (0.92-1.0) | 0.95 (0.84-1.0) | 0.79 (0.68-0.87) | 0.66 (0.53-0.77) | 1.0 (1.0-1.0) | 0.68 (0.63-0.73) |
| overall | 0.94 (0.88-0.97) | 0.89 (0.84-0.94) | 0.76 (0.67-0.83) | 0.65(0.56-0.74) | 0.86(0.78-0.91) | 0.69 (0.6-0.78) |

FIG. 9B
(CONT.)

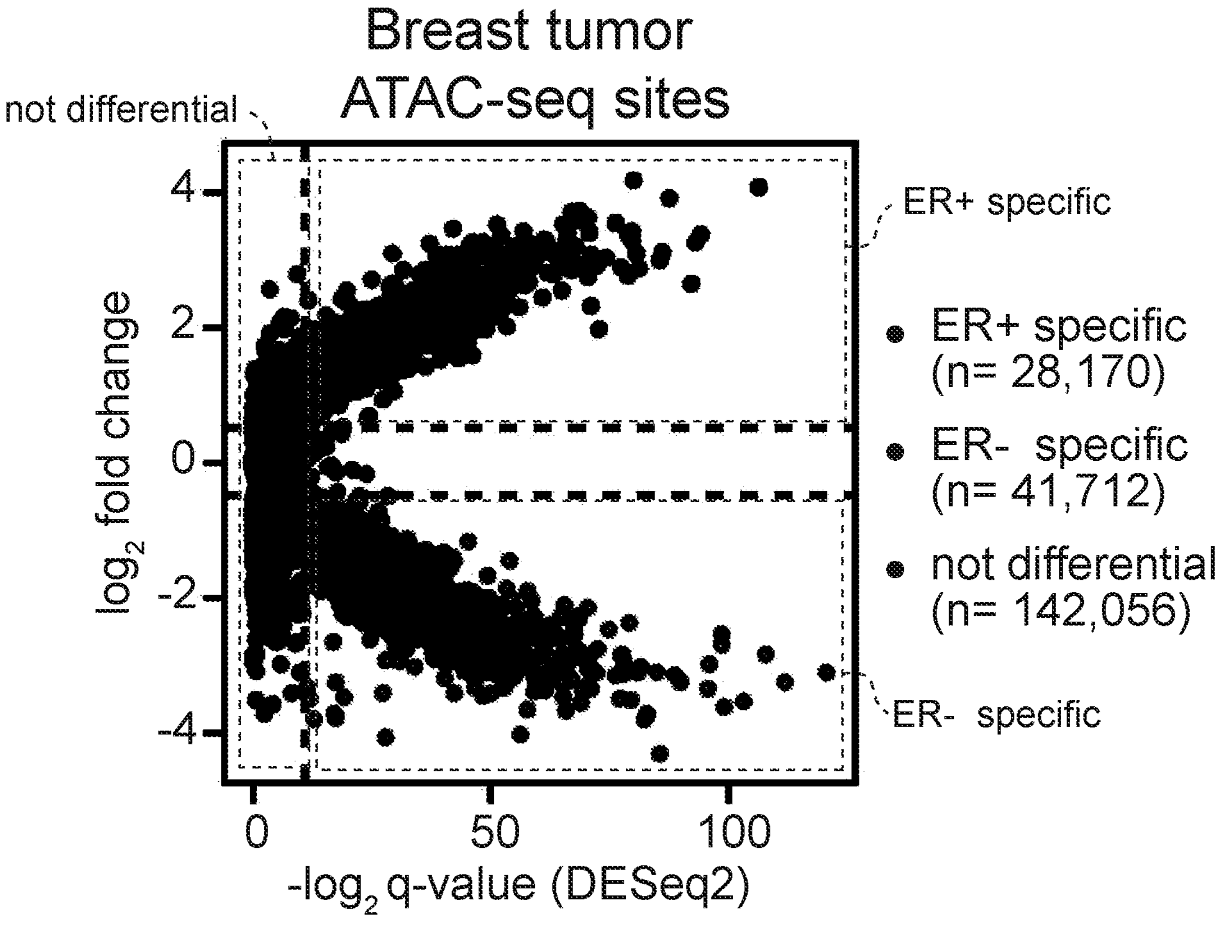

Breast tumor
ATAC-seq sites
not differential
ER+ specific
ER- specific
● ER+ specific (n= 28,170)
● ER- specific (n= 41,712)
● not differential (n= 142,056)
log₂ fold change
4
2
0
-2
-4
0
50
100
-log₂ q-value (DESeq2)

FIG. 10A

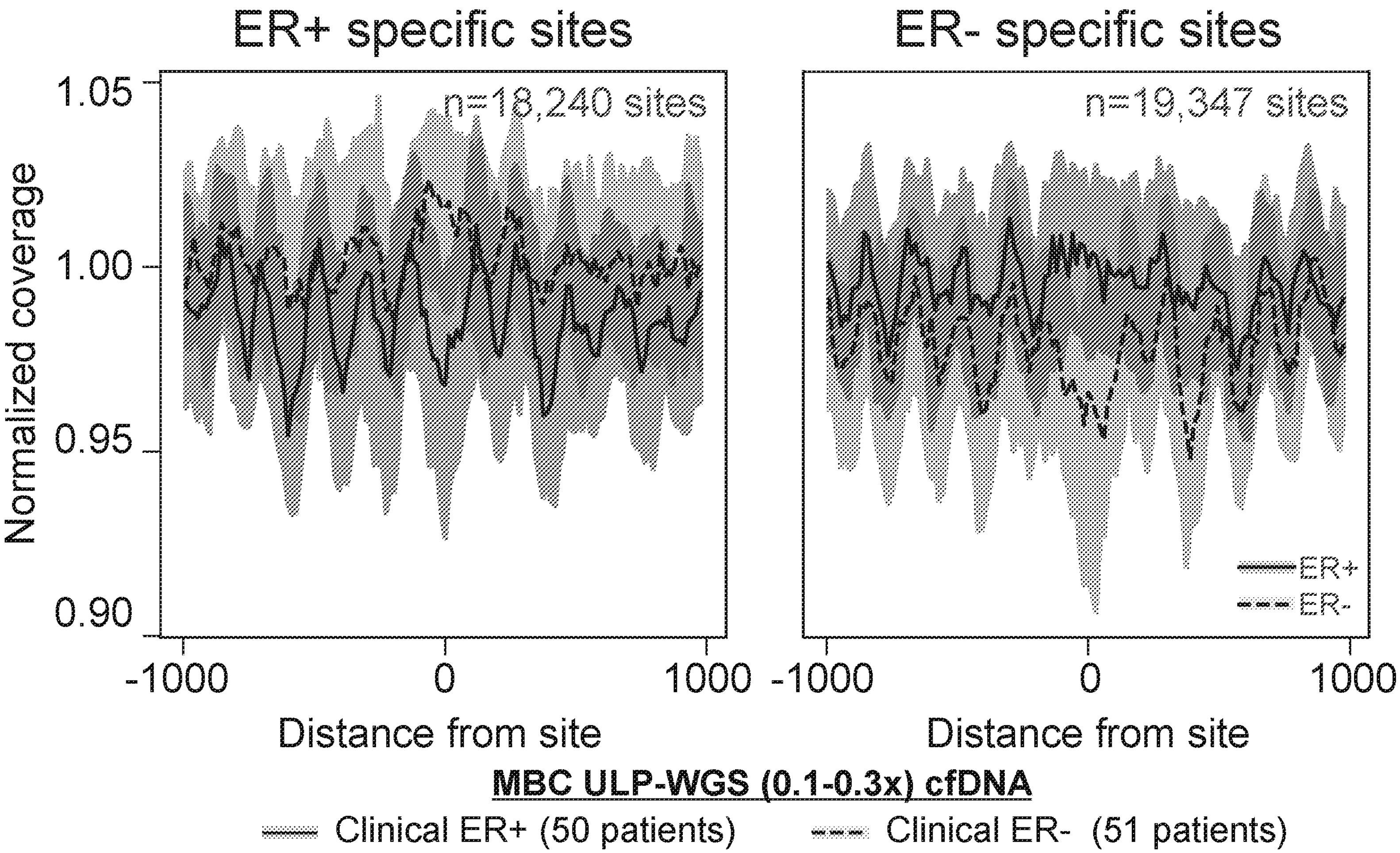

ER+ specific sites
ER- specific sites
n=18,240 sites
n=19,347 sites
Normalized coverage
1.05
1.00
0.95
0.90
-1000
0
1000
-1000
0
1000
Distance from site
Distance from site
ER+
ER-
MBC ULP-WGS (0.1-0.3x) cfDNA
Clinical ER+ (50 patients)
Clinical ER- (51 patients)

FIG. 10B

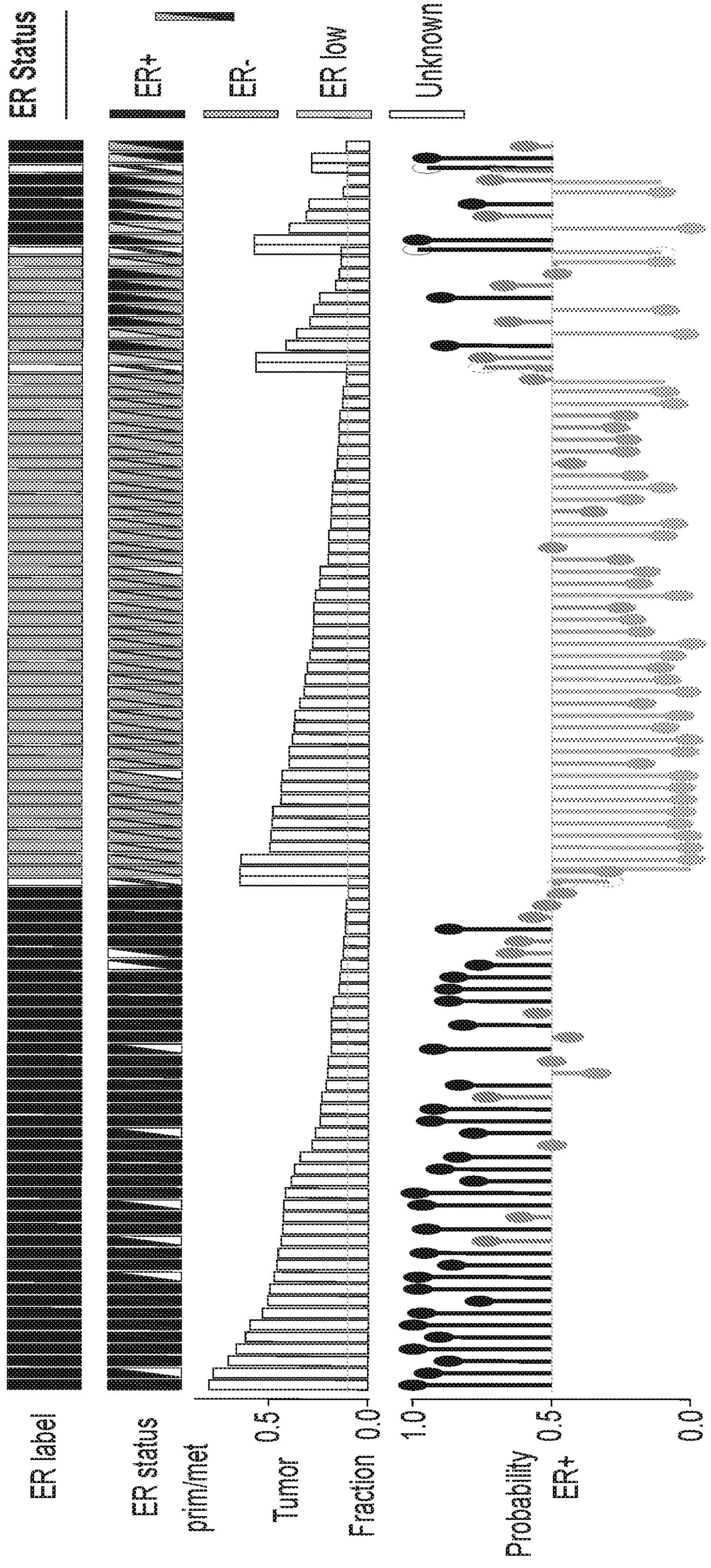

ER Status
ER+
ER-
ER low
Unknown
ER label
ER status
prim/met
Tumor
Fraction
0.5
0.0
1.0
0.5
0.0
Probability
ER+

FIG. 10C

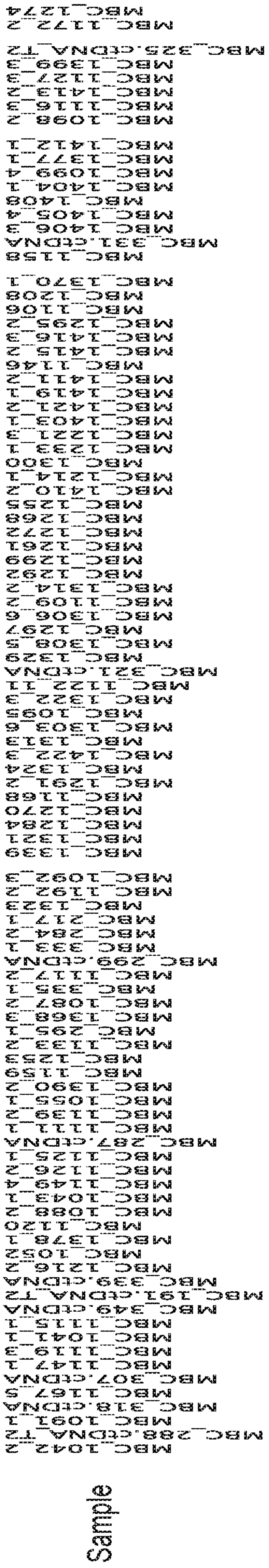
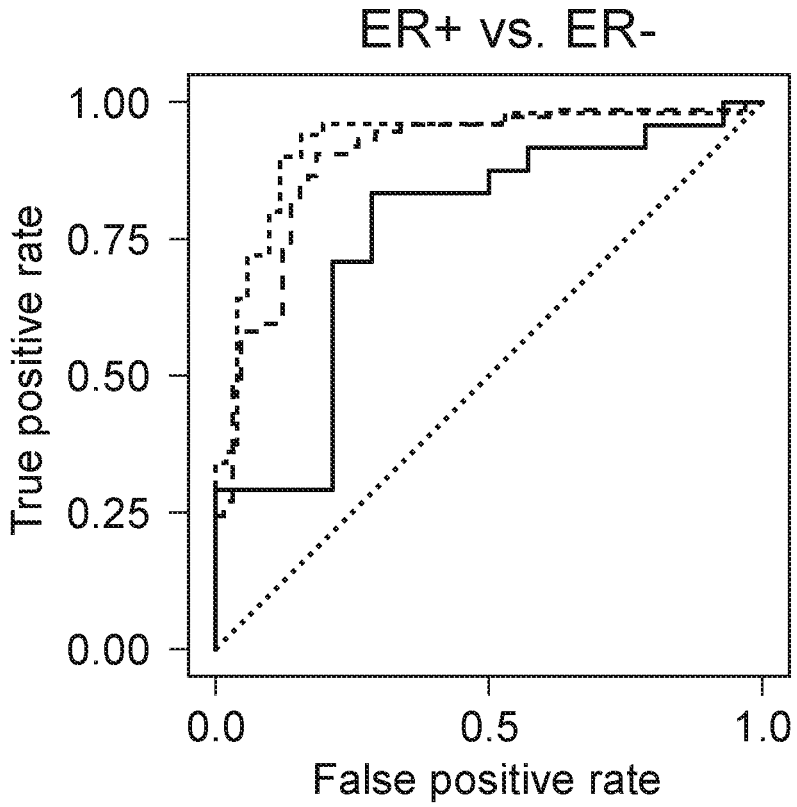
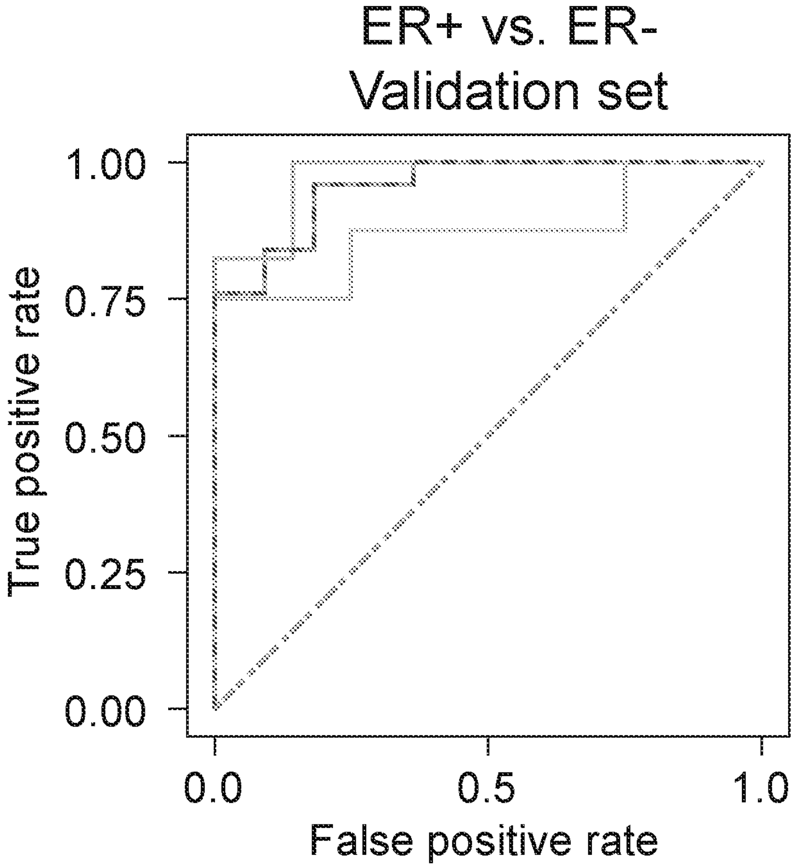

MBC_1274
MBC_1172_2
MBC_325.ctDNA_T2
MBC_1399_3
MBC_1127_3
MBC_1413_2
MBC_1116_3
MBC_1098_2
MBC_1412_1
MBC_1377_1
MBC_1099_4
MBC_1404_1
MBC_1408
MBC_1405_4
MBC_1406_3
MBC_331.ctDNA
MBC_1158
MBC_1370_1
MBC_1208
MBC_1106
MBC_1295_2
MBC_1416_3
MBC_1415_2
MBC_1146
MBC_1411_2
MBC_1419_1
MBC_1421_2
MBC_1403_1
MBC_1221_3
MBC_1233_1
MBC_1300
MBC_1214_1
MBC_1410_2
MBC_1255
MBC_1268
MBC_1272
MBC_1261
MBC_1299
MBC_1292
MBC_1314_2
MBC_1109_2
MBC_1306_6
MBC_1297
MBC_1308_5
MBC_1329
MBC_321.ctDNA
MBC_1122_11
MBC_1322_3
MBC_1095
MBC_1303_6
MBC_1313
MBC_1422_3
MBC_1324
MBC_1291_2
MBC_1168
MBC_1270
MBC_1284
MBC_1321
MBC_1339
MBC_1092_3
MBC_1192_2
MBC_1323
MBC_217_1
MBC_284_2
MBC_333_1
MBC_299.ctDNA
MBC_1117_2
MBC_335_1
MBC_1087_2
MBC_1368_3
MBC_295_1
MBC_1133_2
MBC_1253
MBC_1159
MBC_1390_2
MBC_1055_1
MBC_1139_2
MBC_1111_1
MBC_287.ctDNA
MBC_1125_1
MBC_1126_2
MBC_1149_4
MBC_1043_1
MBC_1088_2
MBC_1120
MBC_1378_1
MBC_1052
MBC_1216_2
MBC_339.ctDNA
MBC_191.ctDNA_T2
MBC_349.ctDNA
MBC_1115_1
MBC_1041_1
MBC_1119_3
MBC_1147_1
MBC_307.ctDNA
MBC_1167_5
MBC_318.ctDNA
MBC_1091_1
MBC_288.ctDNA_T2
MBC_1042_2
Sample

FIG. 10C (CONT.)



| | TFx | n | Accuracy | AUC |
|---|---|---|---|---|
| — | 0.05-0.1 | 38 | 0.69 (0.38-0.92) | 0.75 (0.42-1.00) |
| ···· | ≥0.1 | 101 | 0.86 (0.76-0.95) | 0.92 (0.84-0.99) |
| - - - | All >0.05 | 139 | 0.81 (0.70-0.90) | 0.89 (0.81-0.96) |

*FIG. 10D*



| | TFx | n | Accuracy | AUC |
|---|---|---|---|---|
| —— | 0.05-0.1 | 12 | 0.85 (0.60-1.00) | 0.90 (0.60-1.00) |
| ---- | ≥0.1 | 24 | 0.96 (0.86-1.00) | 0.98 (0.89-1.00) |
| - - · | All >0.05 | 36 | 0.92 (0.82-1.00) | 0.96 (0.88-1.00) |

*FIG. 10E*

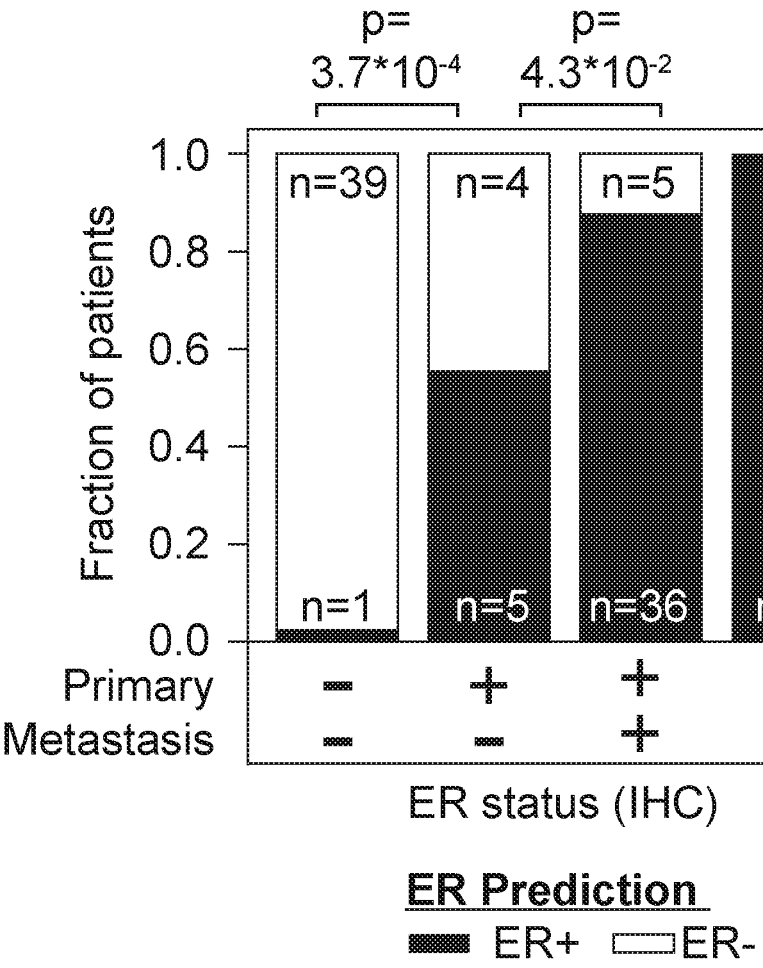
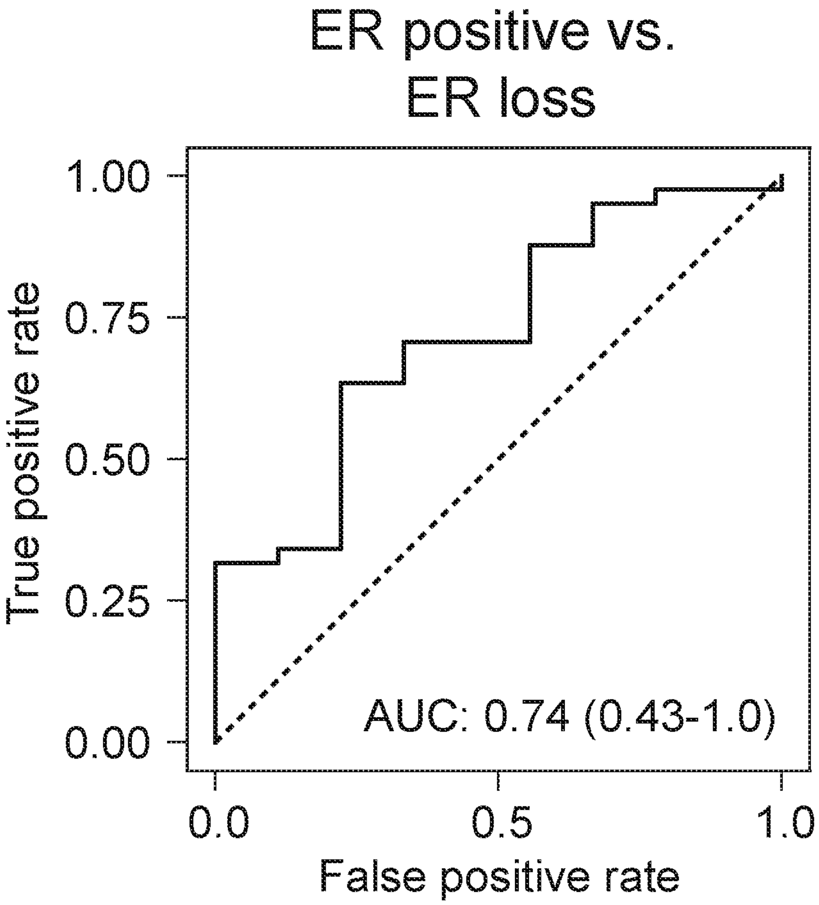

p=
3.7*10-4
p=
4.3*10-2
1.0
0.8
0.6
0.4
0.2
0.0
Fraction of patients
n=39
n=1
n=4
n=5
n=5
n=36
n=1
Primary
Metastasis
-
-
+
-
+
+
-
+
ER status (IHC)
ER Prediction
ER+
ER-

*FIG. 10F*



| Primary Biopsy (IHC) | Metastatic Biopsy (IHC) | n |
|---|---|---|
| ER+ | ER+ | 41 |
| ER+ | ER- | 9 |

*FIG. 10G*

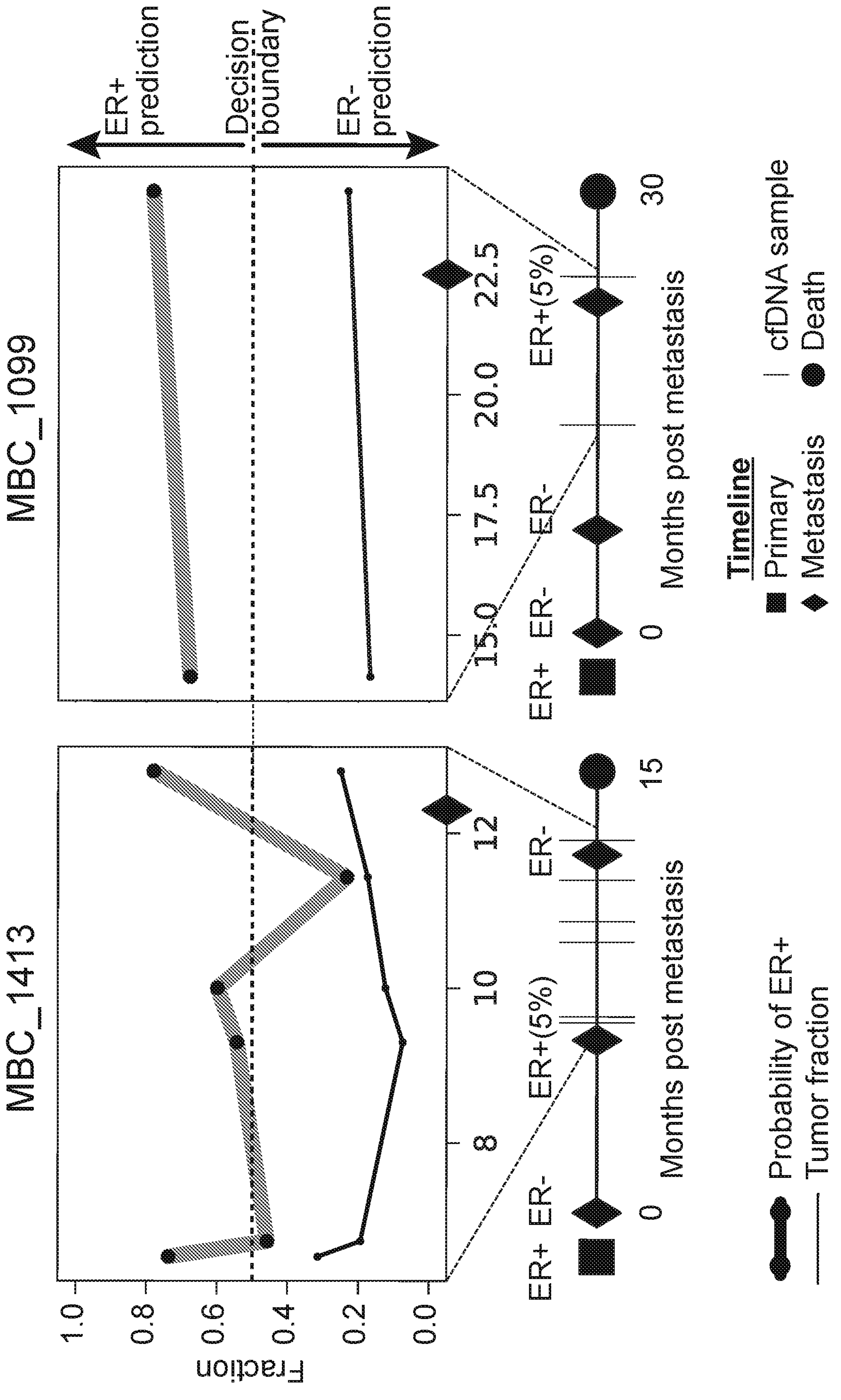

MBC_1413
MBC_1099
Fraction
1.0
0.8
0.6
0.4
0.2
0.0
8
10
12
15.0
17.5
20.0
22.5
ER+ prediction
Decision boundary
ER- prediction
ER+
ER-
ER+(5%)
ER-
ER+
ER-
ER-
ER+(5%)
0
15
0
30
Months post metastasis
Months post metastasis
Probability of ER+
Tumor fraction
Timeline
Primary
Metastasis
cfDNA sample
Death

FIG. 10H

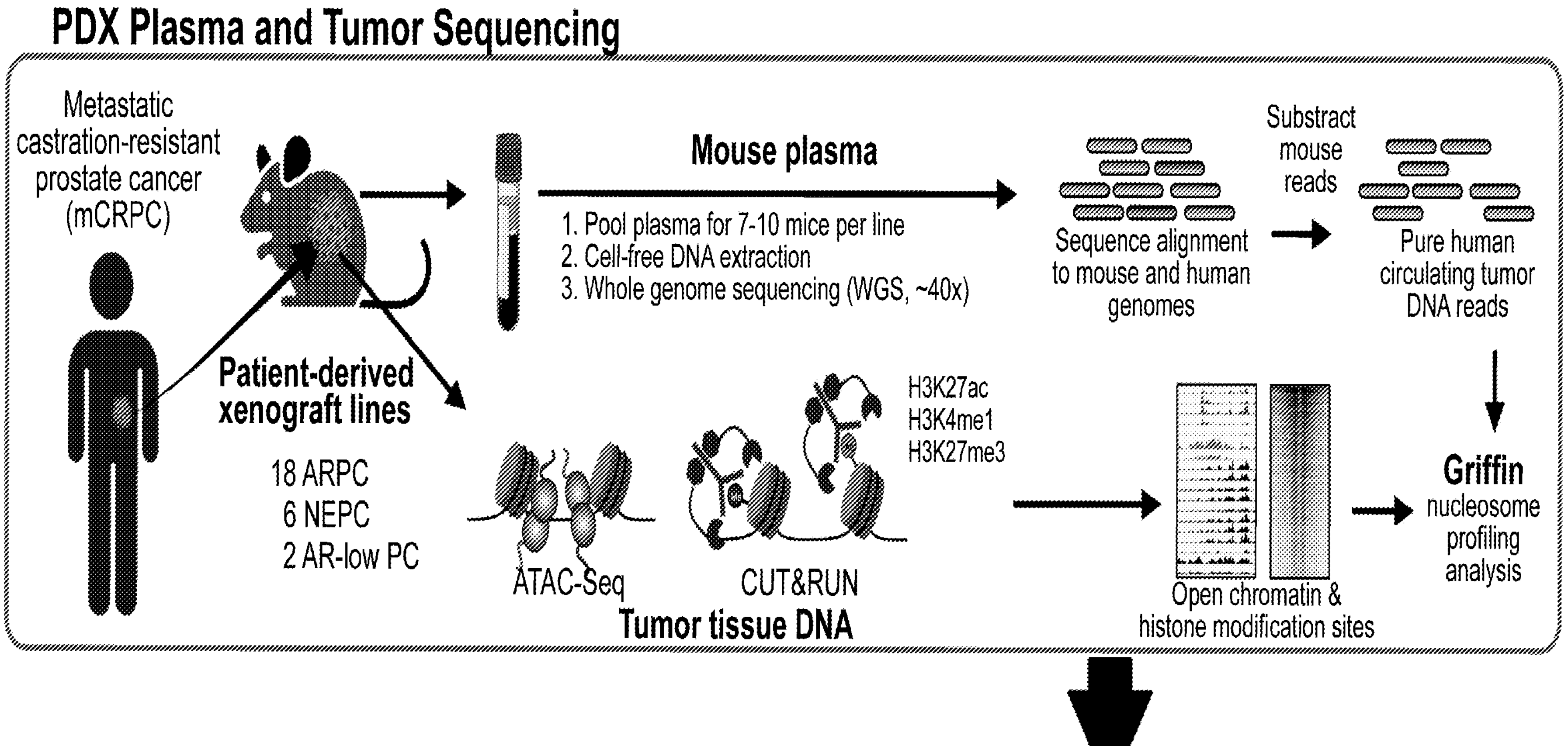

PDX Plasma and Tumor Sequencing
Metastatic castration-resistant prostate cancer (mCRPC)
Patient-derived xenograft lines
18 ARPC
6 NEPC
2 AR-low PC
Mouse plasma
1. Pool plasma for 7-10 mice per line
2. Cell-free DNA extraction
3. Whole genome sequencing (WGS, ~40x)
Sequence alignment to mouse and human genomes
Substract mouse reads
Pure human circulating tumor DNA reads
ATAC-Seq
CUT&RUN
H3K27ac
H3K4me1
H3K27me3
Tumor tissue DNA
Open chromatin & histone modification sites
Griffin
nucleosome profiling analysis

FIG. 11A

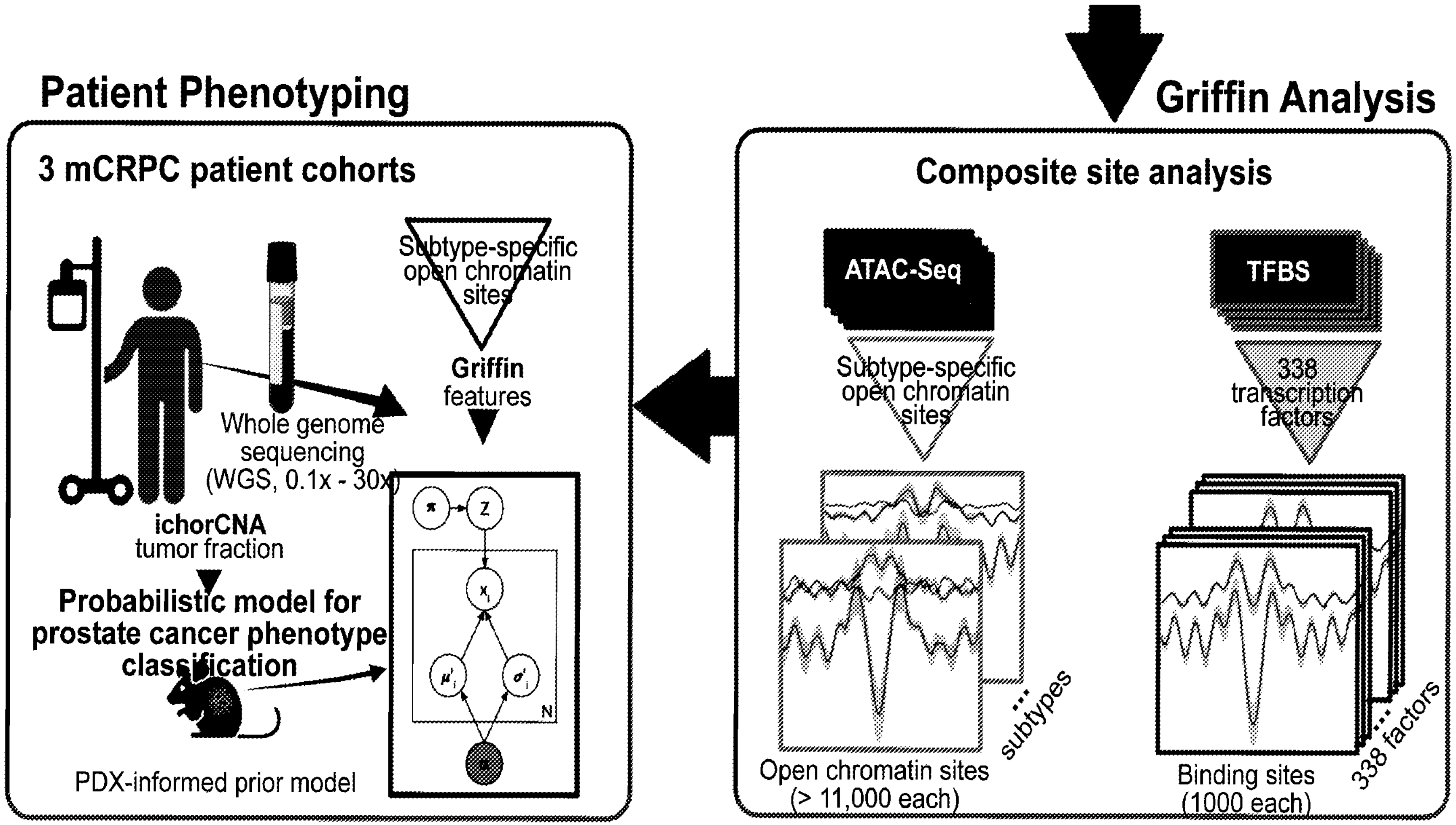

Griffin Analysis
Composite site analysis
ATAC-Seq
Subtype-specific open chromatin sites
Open chromatin sites (> 11,000 each)
subtypes
TFBS
338 transcription factors
Binding sites (1000 each)
338 factors
Patient Phenotyping
3 mCRPC patient cohorts
Whole genome sequencing (WGS, 0.1x - 30x)
ichorCNA tumor fraction
Subtype-specific open chromatin sites
Griffin features
Probabilistic model for prostate cancer phenotype classification
PDX-informed prior model
π
Z
X_i
μ'_i
σ'_i
N

FIG. 11A
(CONT.)

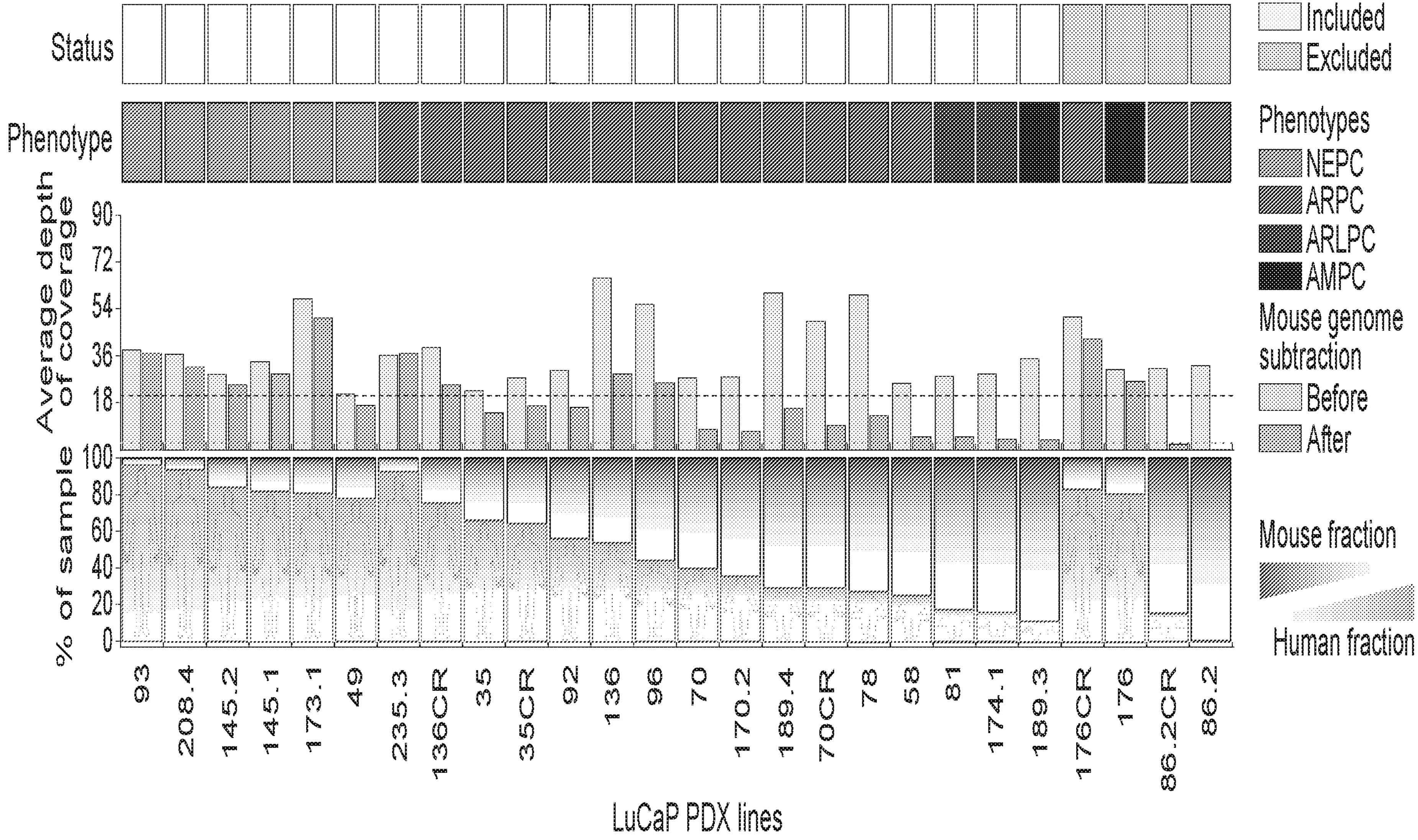

Status
Phenotype
Average depth of coverage
% of sample
90
72
54
36
18
100
80
60
40
20
0
93
208.4
145.2
145.1
173.1
49
235.3
136CR
35
35CR
92
136
96
70
170.2
189.4
70CR
78
58
81
174.1
189.3
176CR
176
86.2CR
86.2
LuCaP PDX lines
Included
Excluded
Phenotypes
NEPC
ARPC
ARLPC
AMPC
Mouse genome subtraction
Before
After
Mouse fraction
Human fraction

FIG. 11B

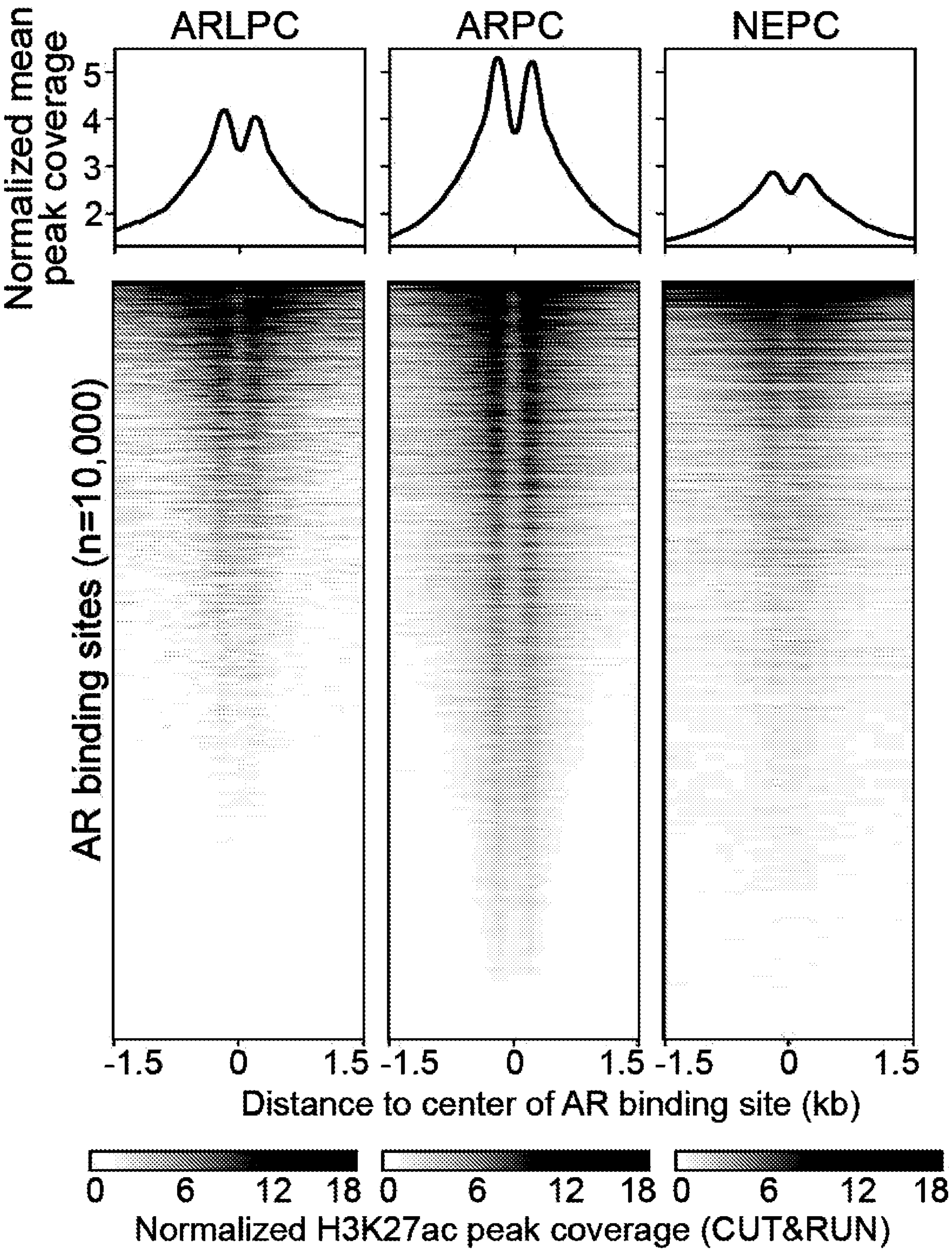

ARLPC
ARPC
NEPC
Normalized mean peak coverage
5
4
3
2
AR binding sites (n=10,000)
-1.5
0
1.5
Distance to center of AR binding site (kb)
0
6
12
18
Normalized H3K27ac peak coverage (CUT&RUN)

FIG. 12A

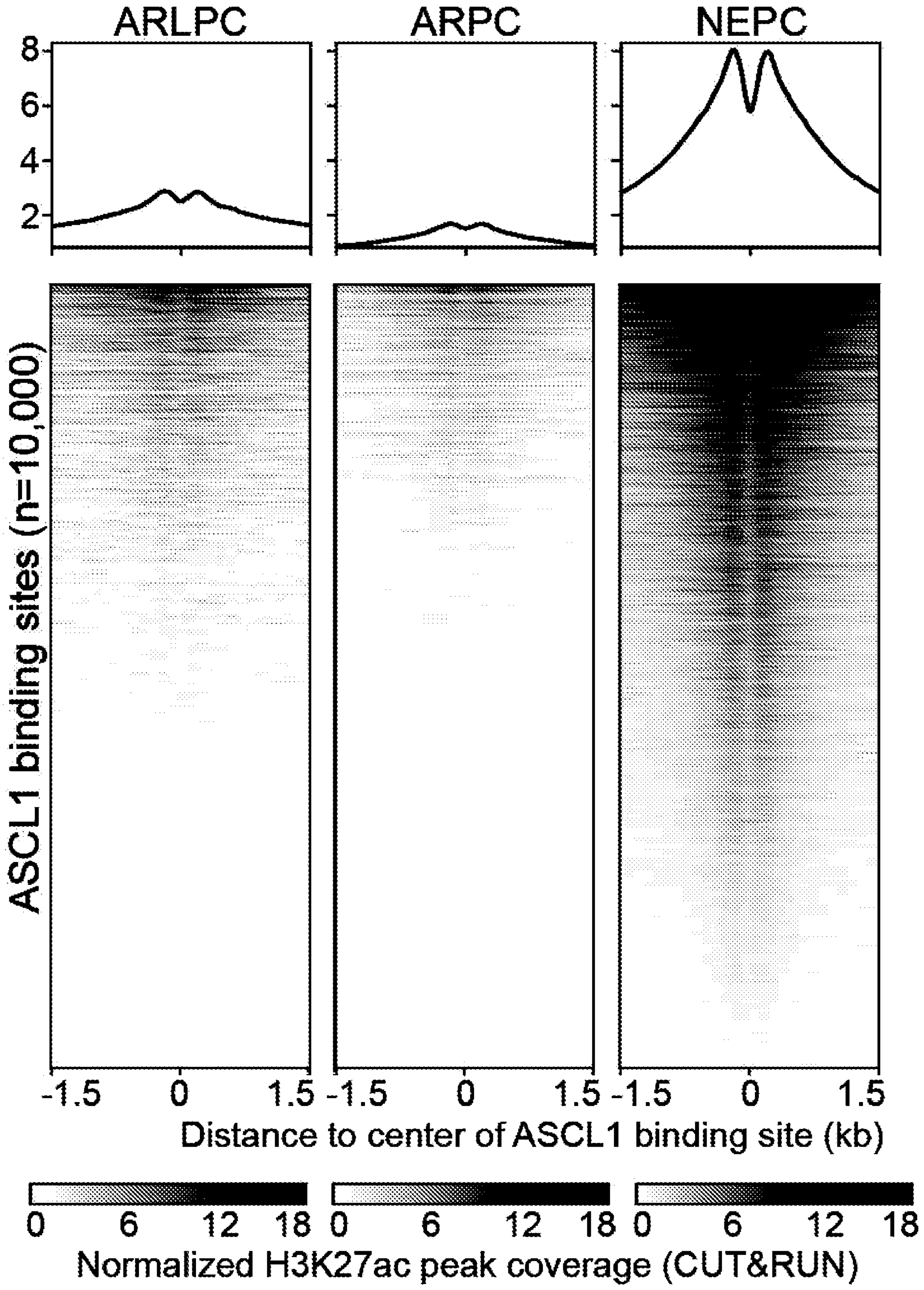

ARLPC
ARPC
NEPC
8
6
4
2
ASCL1 binding sites (n=10,000)
-1.5
0
1.5
-1.5
0
1.5
-1.5
0
1.5
Distance to center of ASCL1 binding site (kb)
0
6
12
18
0
6
12
18
0
6
12
18
Normalized H3K27ac peak coverage (CUT&RUN)

*FIG. 12A*
*(CONT.)*

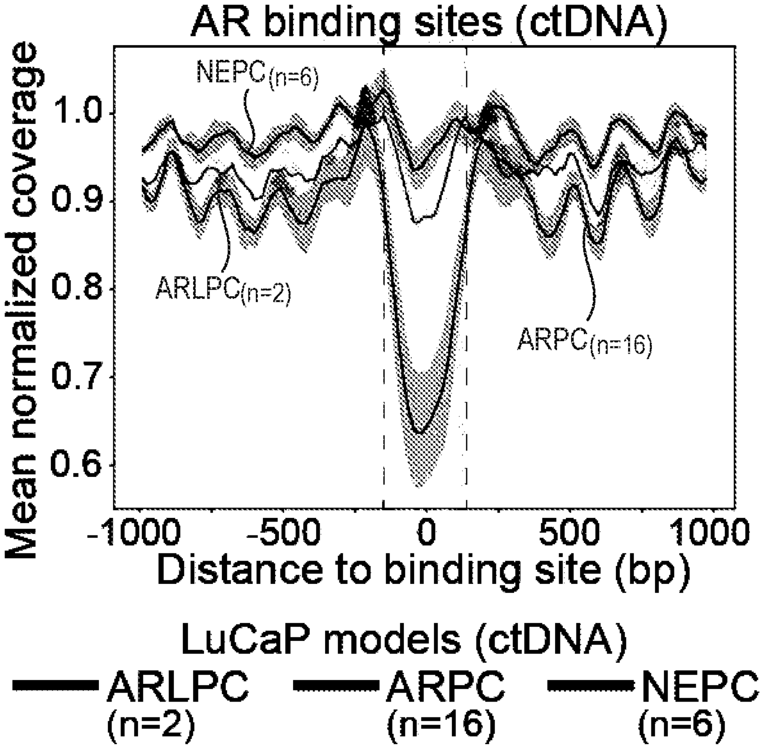


FIG. 12B

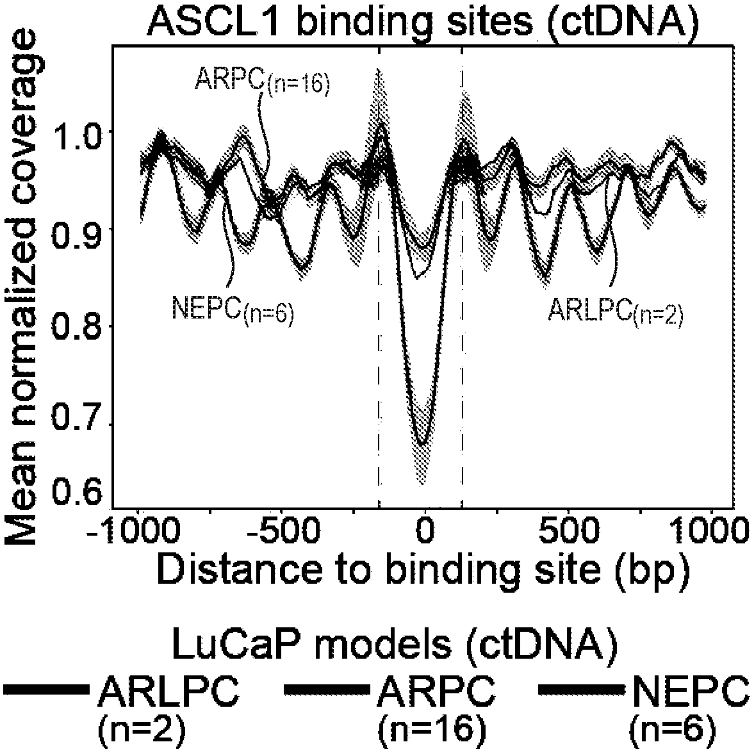


FIG. 12C

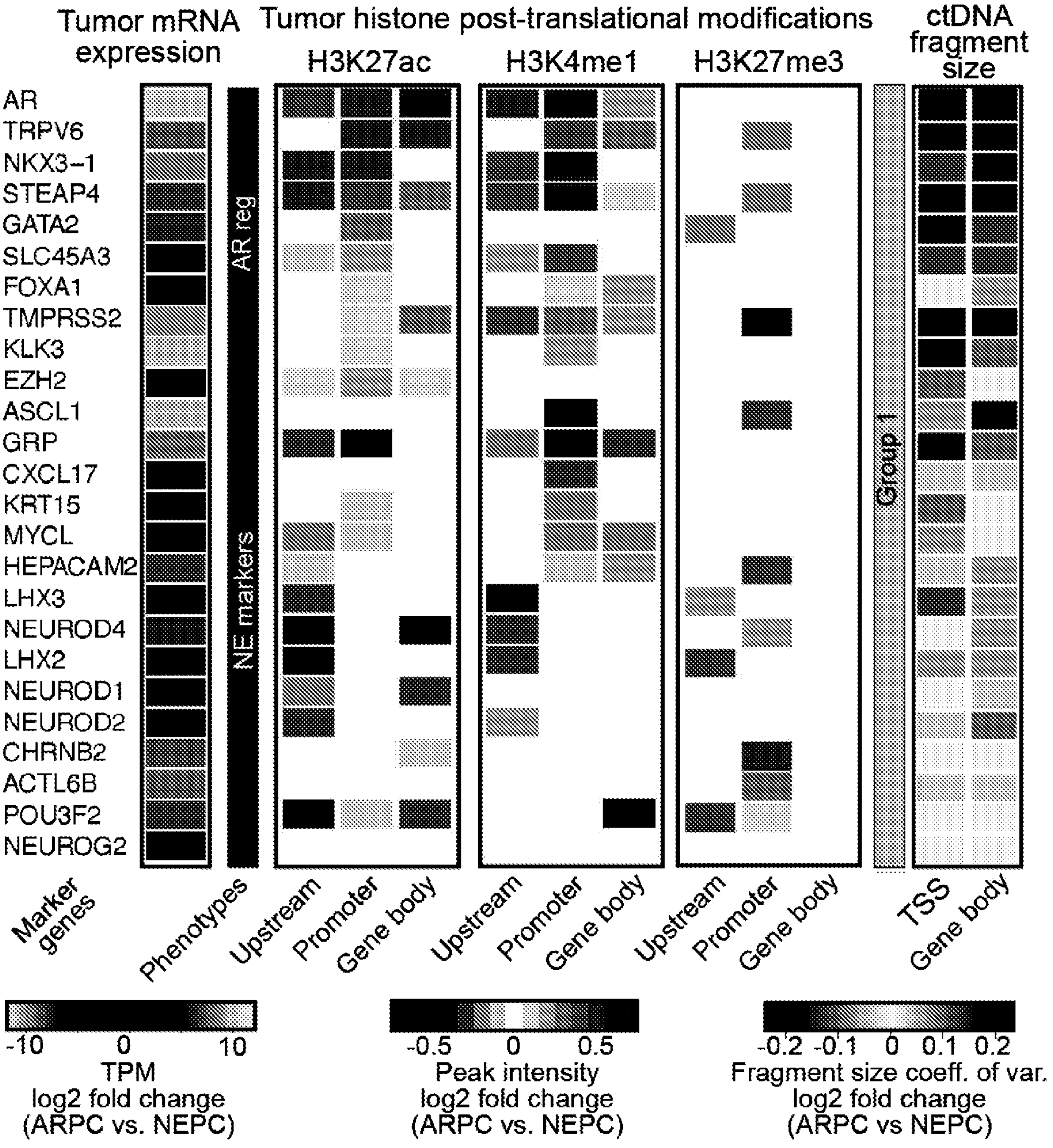

Tumor mRNA expression
Tumor histone post-translational modifications
H3K27ac
H3K4me1
H3K27me3
ctDNA fragment size
AR
TRPV6
NKX3-1
STEAP4
GATA2
SLC45A3
FOXA1
TMPRSS2
KLK3
EZH2
ASCL1
GRP
CXCL17
KRT15
MYCL
HEPACAM2
LHX3
NEUROD4
LHX2
NEUROD1
NEUROD2
CHRNB2
ACTL6B
POU3F2
NEUROG2
AR reg
NE markers
Group 1
Marker genes
Phenotypes
Upstream
Promoter
Gene body
Upstream
Promoter
Gene body
Upstream
Promoter
Gene body
TSS
Gene body
-10 0 10
TPM
log2 fold change
(ARPC vs. NEPC)
-0.5 0 0.5
Peak intensity
log2 fold change
(ARPC vs. NEPC)
-0.2 -0.1 0 0.1 0.2
Fragment size coeff. of var.
log2 fold change
(ARPC vs NEPC)

*FIG. 12D*

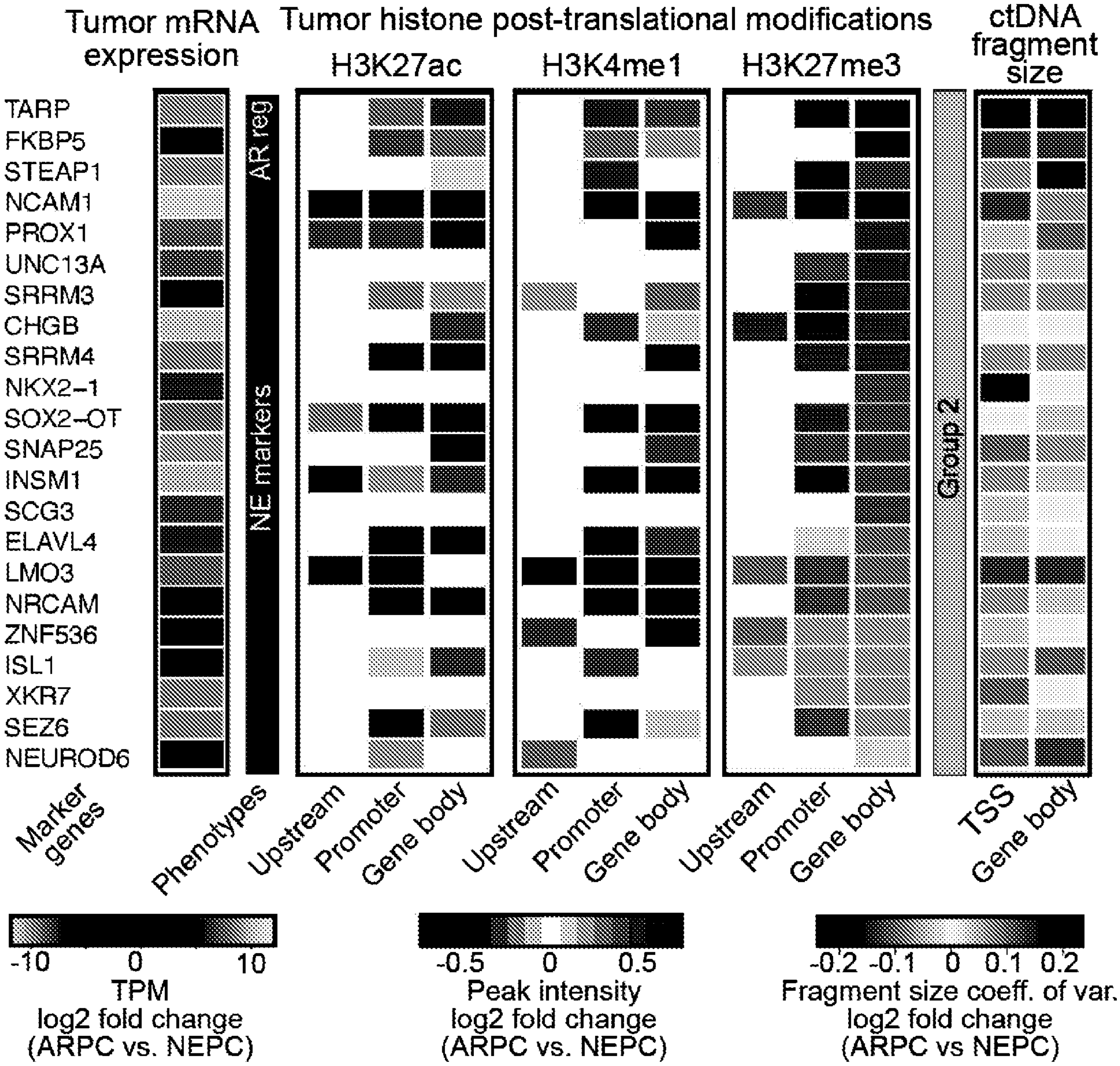

Tumor mRNA expression
Tumor histone post-translational modifications
ctDNA fragment size
H3K27ac
H3K4me1
H3K27me3
TARP
FKBP5
STEAP1
NCAM1
PROX1
UNC13A
SRRM3
CHGB
SRRM4
NKX2-1
SOX2-OT
SNAP25
INSM1
SCG3
ELAVL4
LMO3
NRCAM
ZNF536
ISL1
XKR7
SEZ6
NEUROD6
AR reg
NE markers
Group 2
Marker genes
Phenotypes
Upstream
Promoter
Gene body
Upstream
Promoter
Gene body
Upstream
Promoter
Gene body
TSS
Gene body
-10
0
10
TPM log2 fold change (ARPC vs. NEPC)
-0.5
0
0.5
Peak intensity log2 fold change (ARPC vs. NEPC)
-0.2
-0.1
0
0.1
0.2
Fragment size coeff. of var. log2 fold change (ARPC vs NEPC)

FIG. 12D (CONT.)

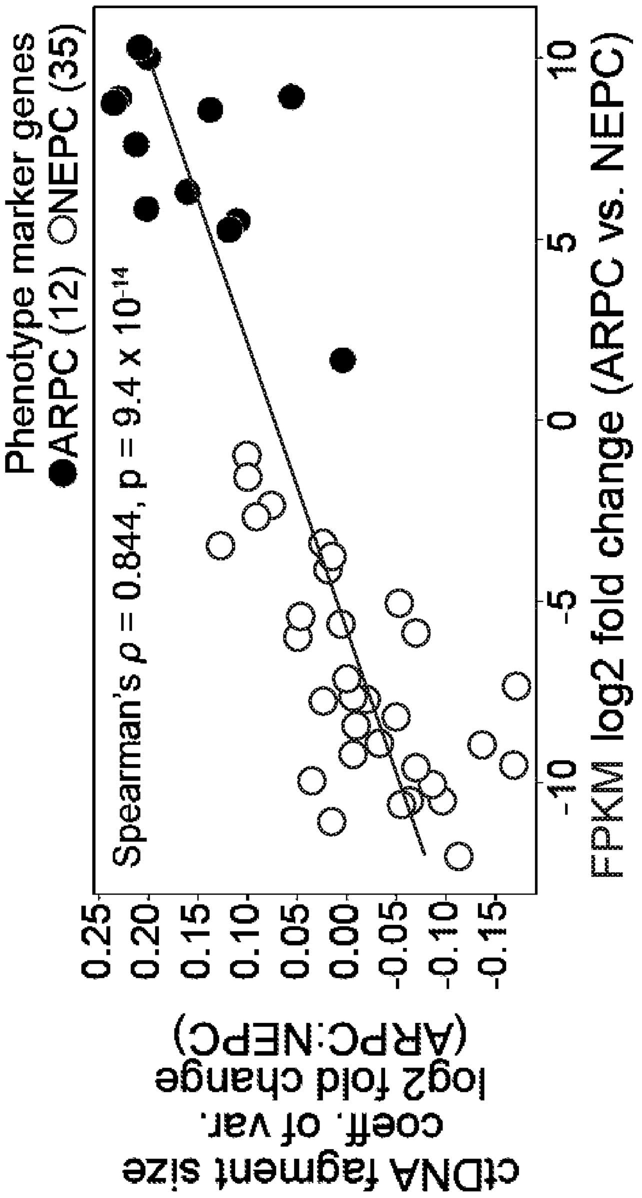

Phenotype marker genes
● ARPC (12) ○ NEPC (35)
Spearman's $\rho$ = 0.844, p = 9.4 x $10^{-14}$
ctDNA fagment size
coeff. of var.
log2 fold change
(ARPC:NEPC)
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-10
-5
0
5
10
FPKM log2 fold change (ARPC vs. NEPC)

*FIG. 12E*

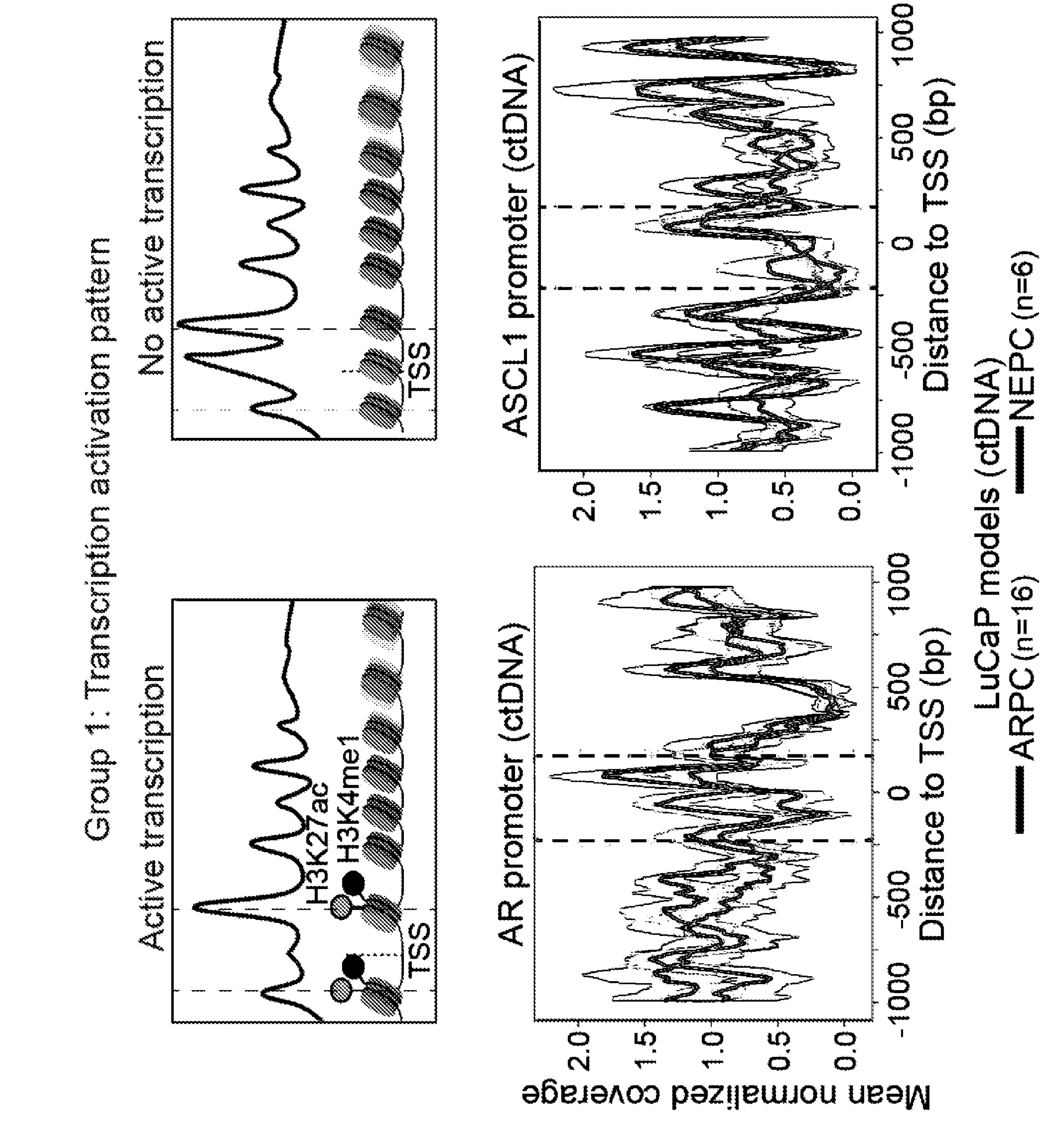

Group 1: Transcription activation pattern
Active transcription
H3K27ac
H3K4me1
TSS
No active transcription
TSS
AR promoter (ctDNA)
ASCL1 promoter (ctDNA)
Mean normalized coverage
2.0
1.5
1.0
0.5
0.0
-1000
-500
0
500
1000
Distance to TSS (bp)
LuCaP models (ctDNA)
ARPC (n=16)
NEPC (n=6)

FIG. 12F

Group 2: Transcription repression pattern

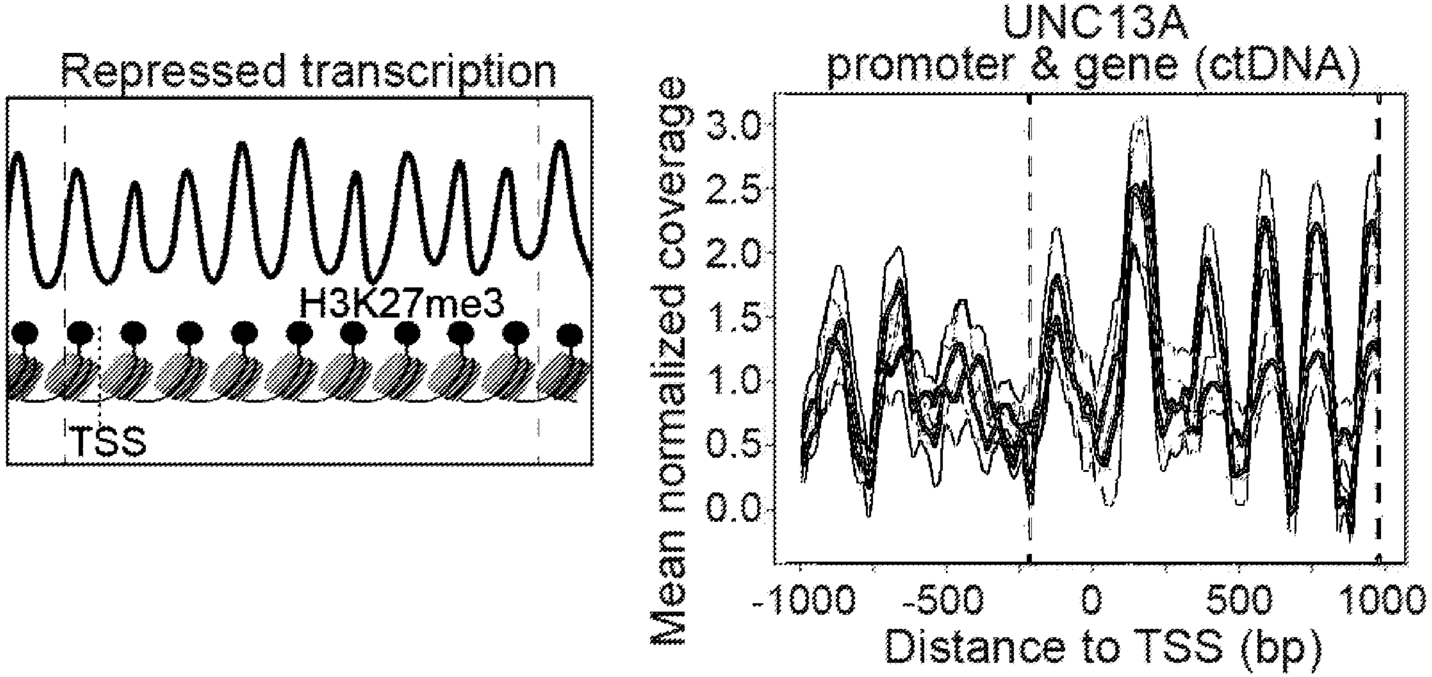


FIG. 12G

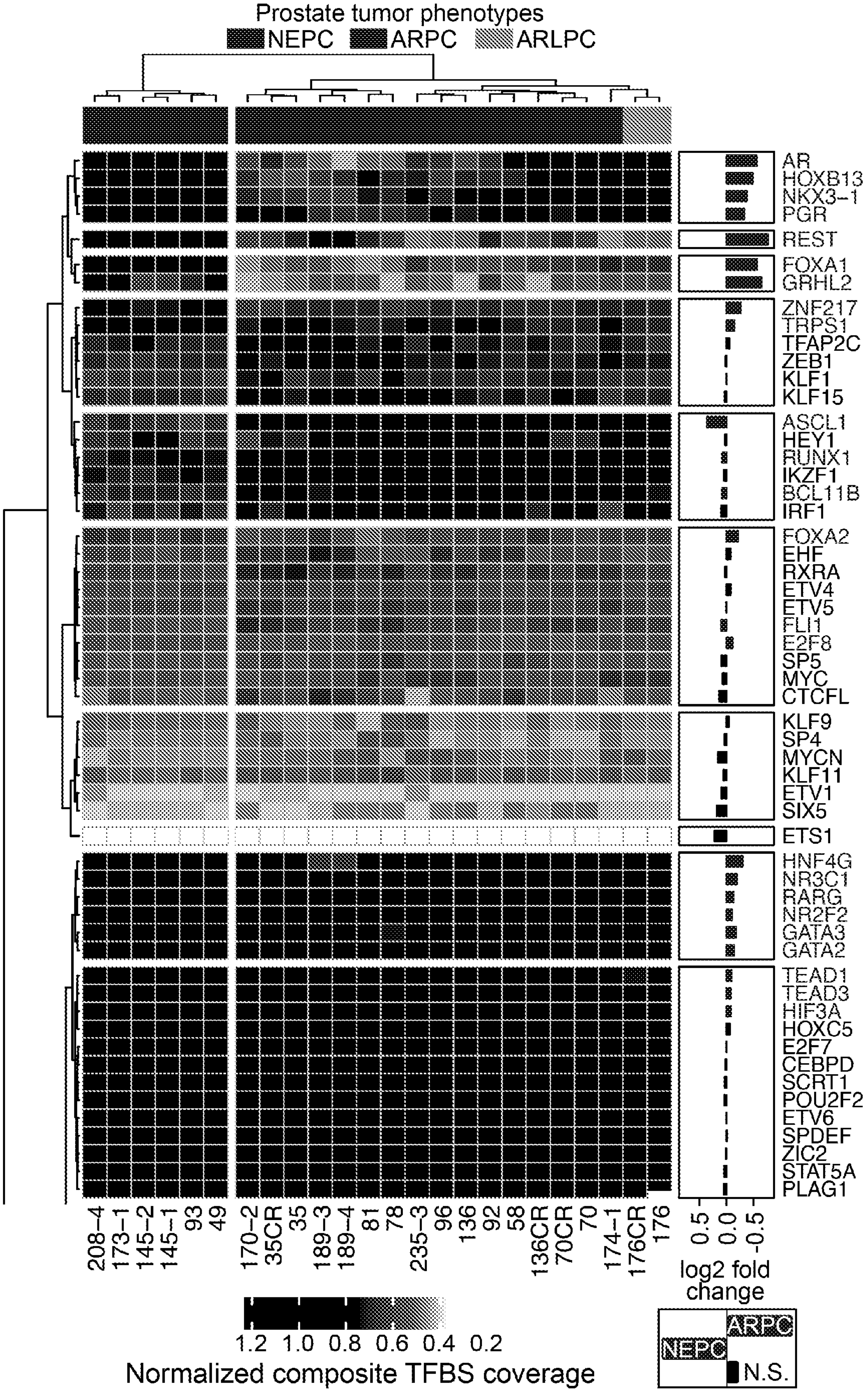

Prostate tumor phenotypes
NEPC
ARPC
ARLPC
AR
HOXB13
NKX3-1
PGR
REST
FOXA1
GRHL2
ZNF217
TRPS1
TFAP2C
ZEB1
KLF1
KLF15
ASCL1
HEY1
RUNX1
IKZF1
BCL11B
IRF1
FOXA2
EHF
RXRA
ETV4
ETV5
FLI1
E2F8
SP5
MYC
CTCFL
KLF9
SP4
MYCN
KLF11
ETV1
SIX5
ETS1
HNF4G
NR3C1
RARG
NR2F2
GATA3
GATA2
TEAD1
TEAD3
HIF3A
HOXC5
E2F7
CEBPD
SCRT1
POU2F2
ETV6
SPDEF
ZIC2
STAT5A
PLAG1
208-4
173-1
145-2
145-1
93
49
170-2
35CR
35
189-3
189-4
81
78
235-3
96
136
92
58
136CR
70CR
70
174-1
176CR
176
0.5
0.0
-0.5
log2 fold change
1.2
1.0
0.8
0.6
0.4
0.2
Normalized composite TFBS coverage
ARPC
NEPC
N.S.

FIG. 13

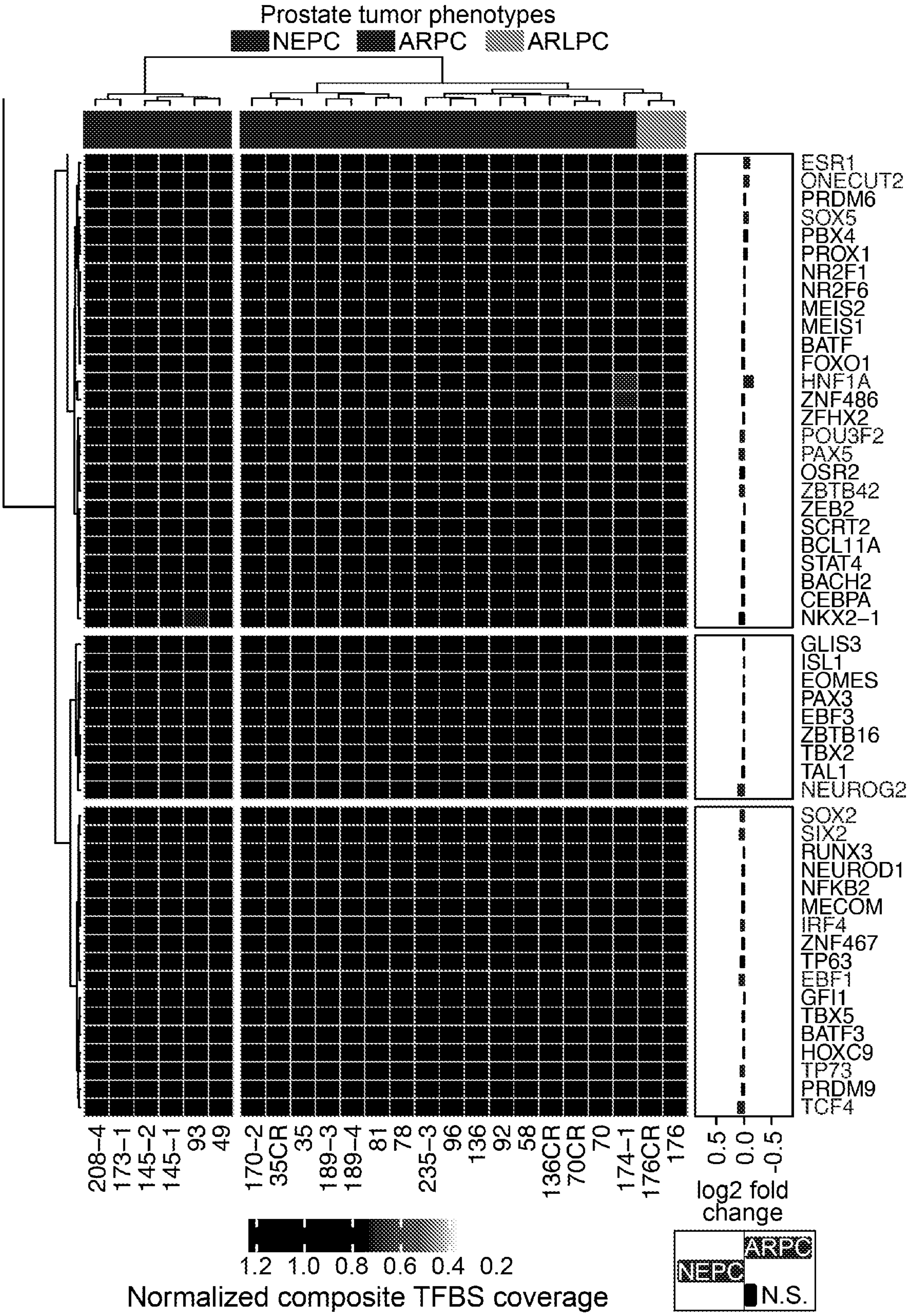
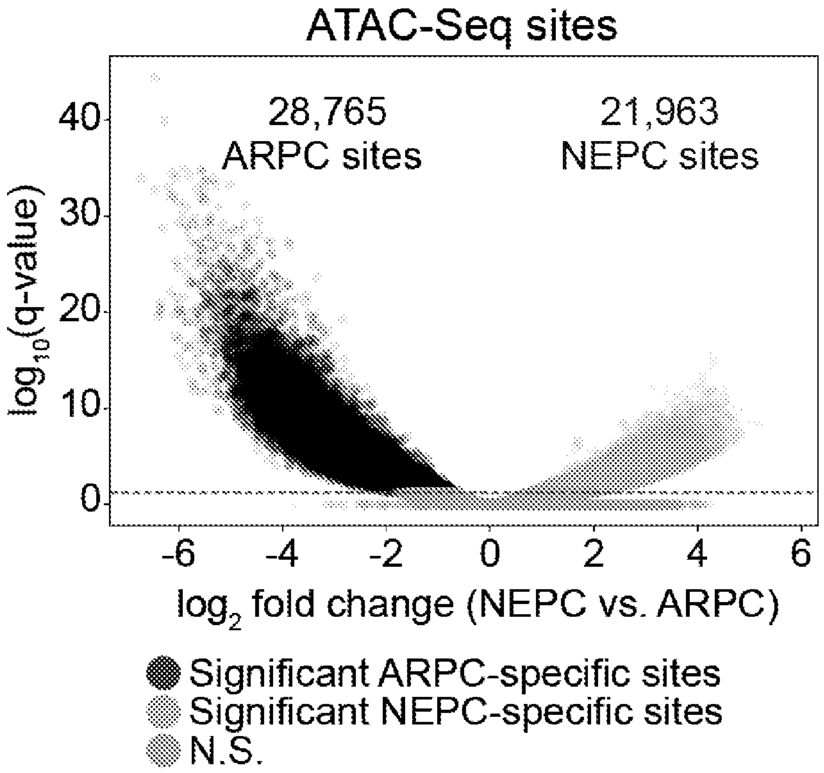
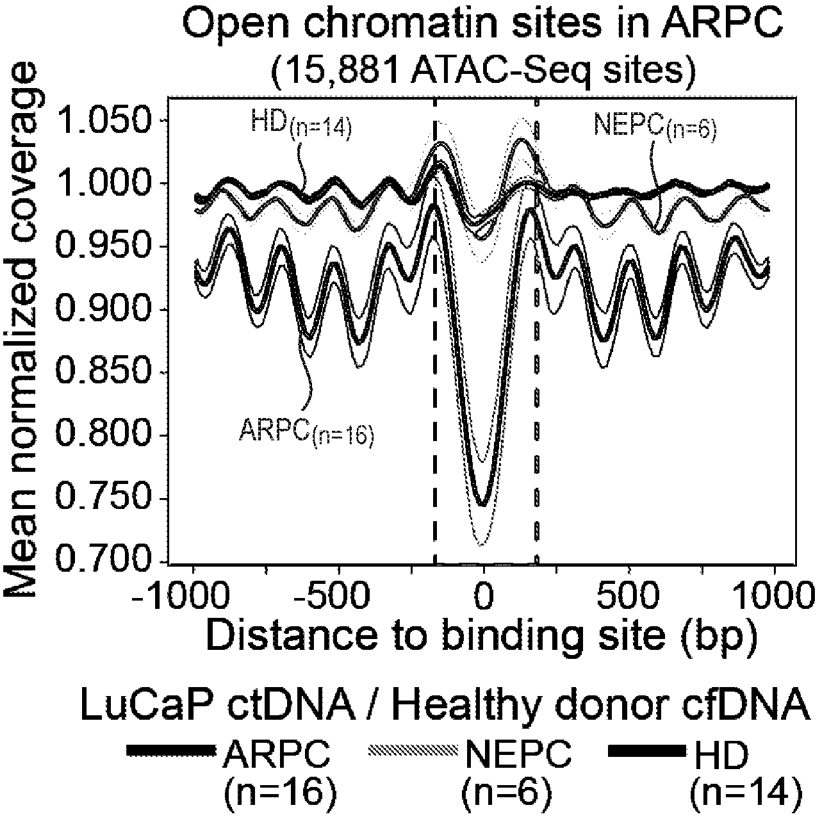

Prostate tumor phenotypes
NEPC
ARPC
ARLPC
ESR1
ONECUT2
PRDM6
SOX5
PBX4
PROX1
NR2F1
NR2F6
MEIS2
MEIS1
BATF
FOXO1
HNF1A
ZNF486
ZFHX2
POU3F2
PAX5
OSR2
ZBTB42
ZEB2
SCRT2
BCL11A
STAT4
BACH2
CEBPA
NKX2–1
GLIS3
ISL1
EOMES
PAX3
EBF3
ZBTB16
TBX2
TAL1
NEUROG2
SOX2
SIX2
RUNX3
NEUROD1
NFKB2
MECOM
IRF4
ZNF467
TP63
EBF1
GFI1
TBX5
BATF3
HOXC9
TP73
PRDM9
TCF4
208–4
173–1
145–2
145–1
93
49
170–2
35CR
35
189–3
189–4
81
78
235–3
96
136
92
58
136CR
70CR
70
174–1
176CR
176
0.5
0.0
-0.5
log2 fold change
1.2 1.0 0.8 0.6 0.4 0.2
Normalized composite TFBS coverage
ARPC
NEPC
N.S.

FIG. 13

(CONT.)



# *FIG. 14A*



# *FIG. 14B*

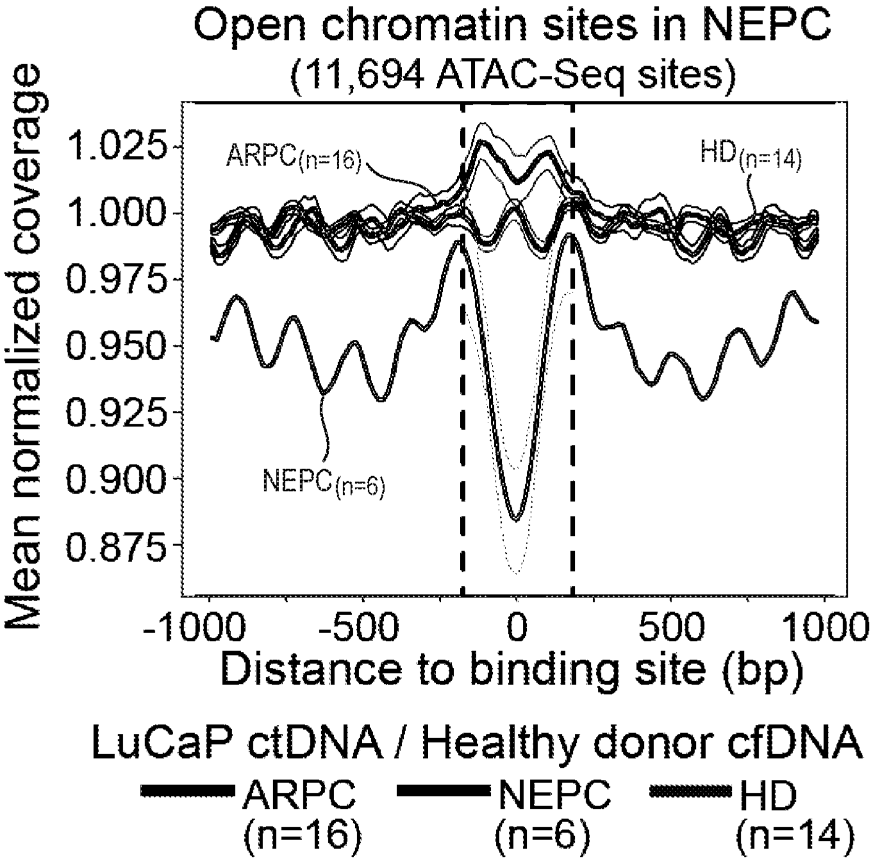

Open chromatin sites in NEPC
(11,694 ATAC-Seq sites)
Mean normalized coverage
1.025
1.000
0.975
0.950
0.925
0.900
0.875
ARPC(n=16)
HD(n=14)
NEPC(n=6)
-1000
-500
0
500
1000
Distance to binding site (bp)
LuCaP ctDNA / Healthy donor cfDNA
ARPC (n=16)
NEPC (n=6)
HD (n=14)

FIG. 14C

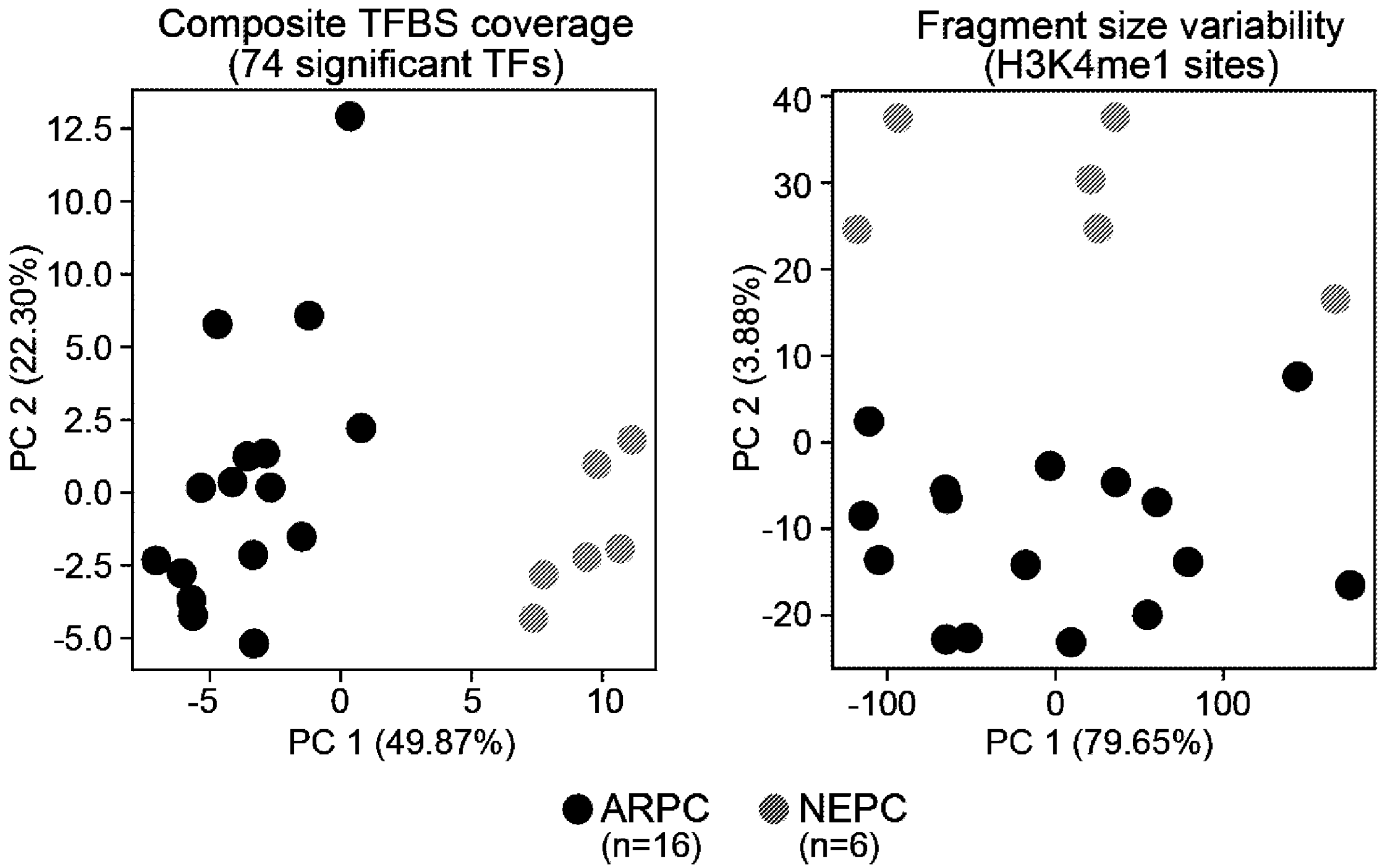

Composite TFBS coverage
(74 significant TFs)
PC 2 (22.30%)
12.5
10.0
10.0
5.0
2.5
0.0
-2.5
-5.0
-5
0
5
10
PC 1 (49.87%)
Fragment size variability
(H3K4me1 sites)
PC 2 (3.88%)
40
30
20
10
0
-10
-20
-100
0
100
PC 1 (79.65%)
● ARPC
(n=16)
▨ NEPC
(n=6)

*FIG. 14D*

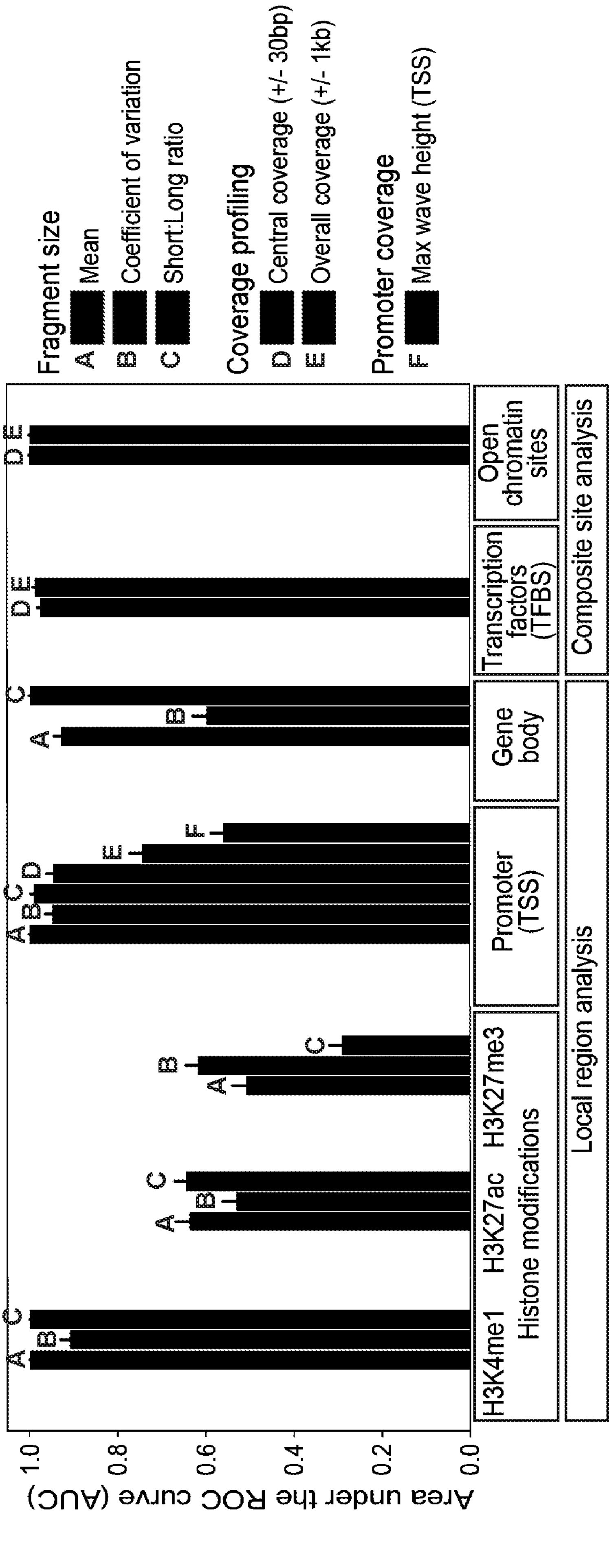

Area under the ROC curve (AUC)
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
E
F
H3K4me1
H3K27ac
H3K27me3
Histone modifications
Promoter (TSS)
Gene body
Transcription factors (TFBS)
Open chromatin sites
Local region analysis
Composite site analysis
Fragment size
A ■ Mean
B ■ Coefficient of variation
C ■ Short:Long ratio
Coverage profiling
D ■ Central coverage (+/- 30bp)
E ■ Overall coverage (+/- 1kb)
Promoter coverage
F ■ Max wave height (TSS)

FIG. 14E

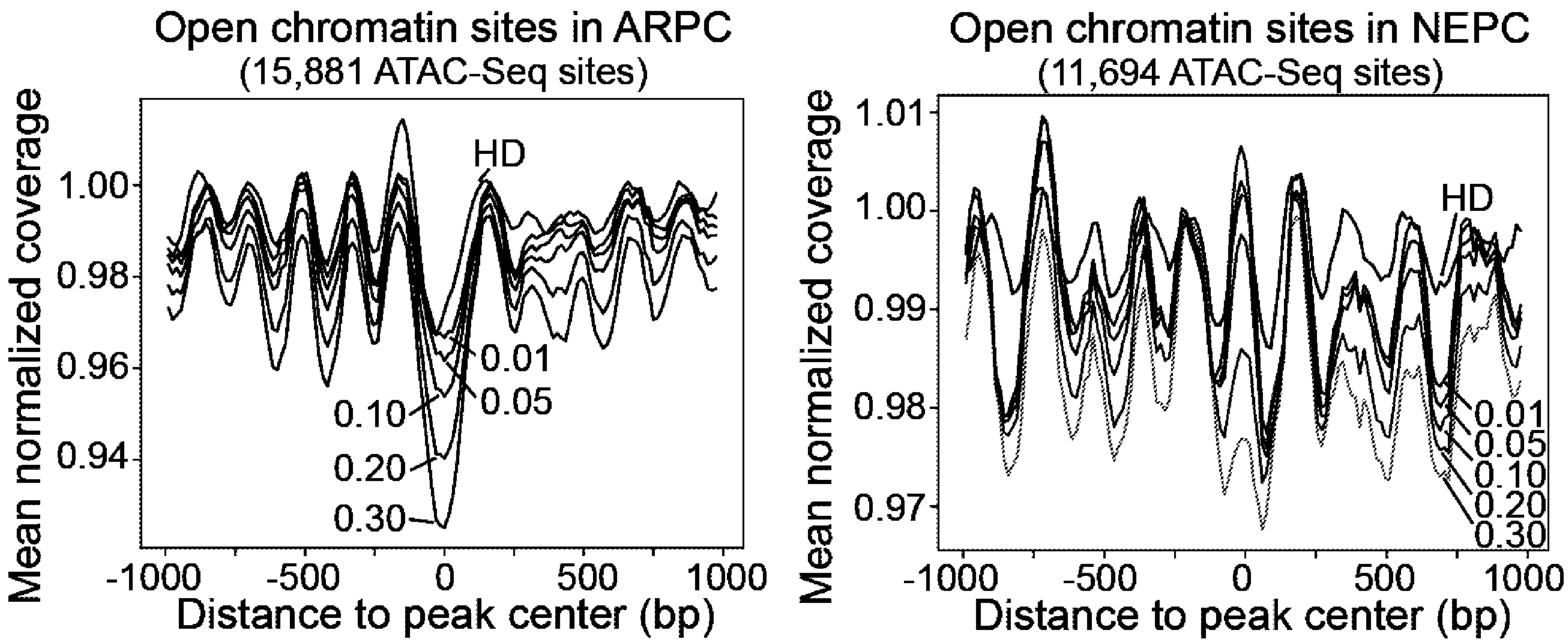

Open chromatin sites in ARPC
(15,881 ATAC-Seq sites)
Mean normalized coverage
1.00
0.98
0.96
0.94
-1000
-500
0
500
1000
Distance to peak center (bp)
HD
0.01
0.05
0.10
0.20
0.30
Open chromatin sites in NEPC
(11,694 ATAC-Seq sites)
Mean normalized coverage
1.01
1.00
0.99
0.98
0.97
-1000
-500
0
500
1000
Distance to peak center (bp)
HD
0.01
0.05
0.10
0.20
0.30

FIG. 14F

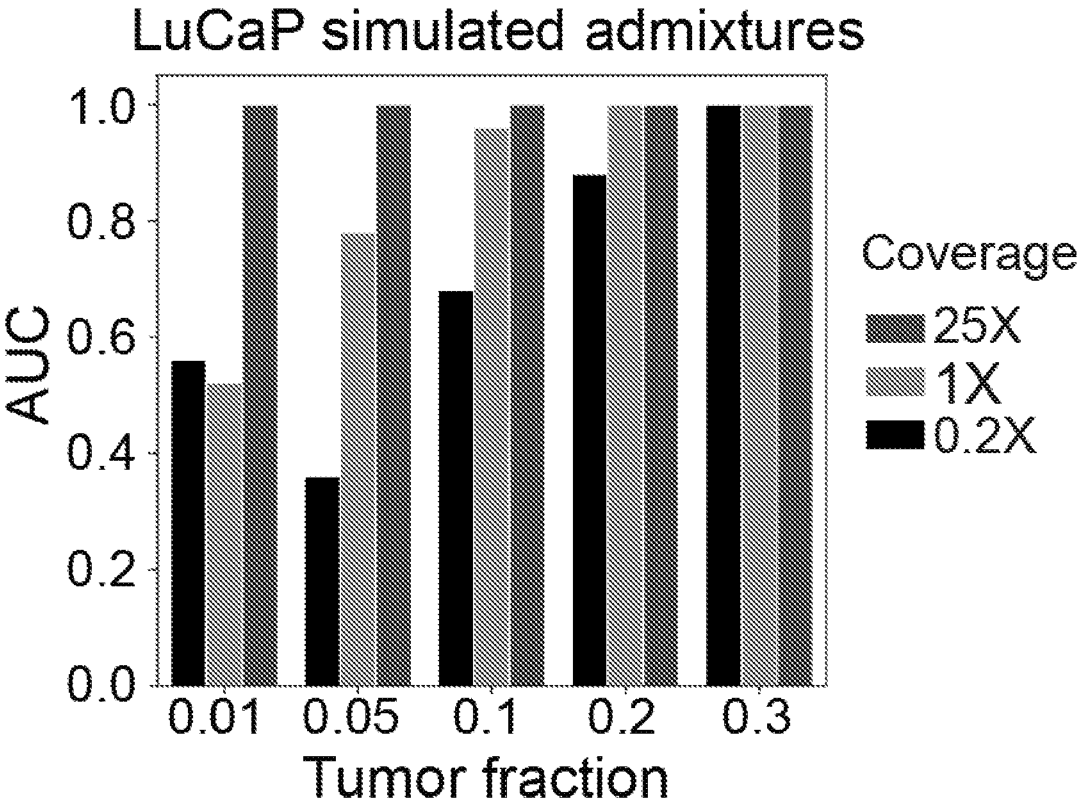

LuCaP simulated admixtures
AUC
1.0
0.8
0.6
0.4
0.2
0.0
0.01
0.05
0.1
0.2
0.3
Tumor fraction
Coverage
25X
1X
0.2X

*FIG. 14G*

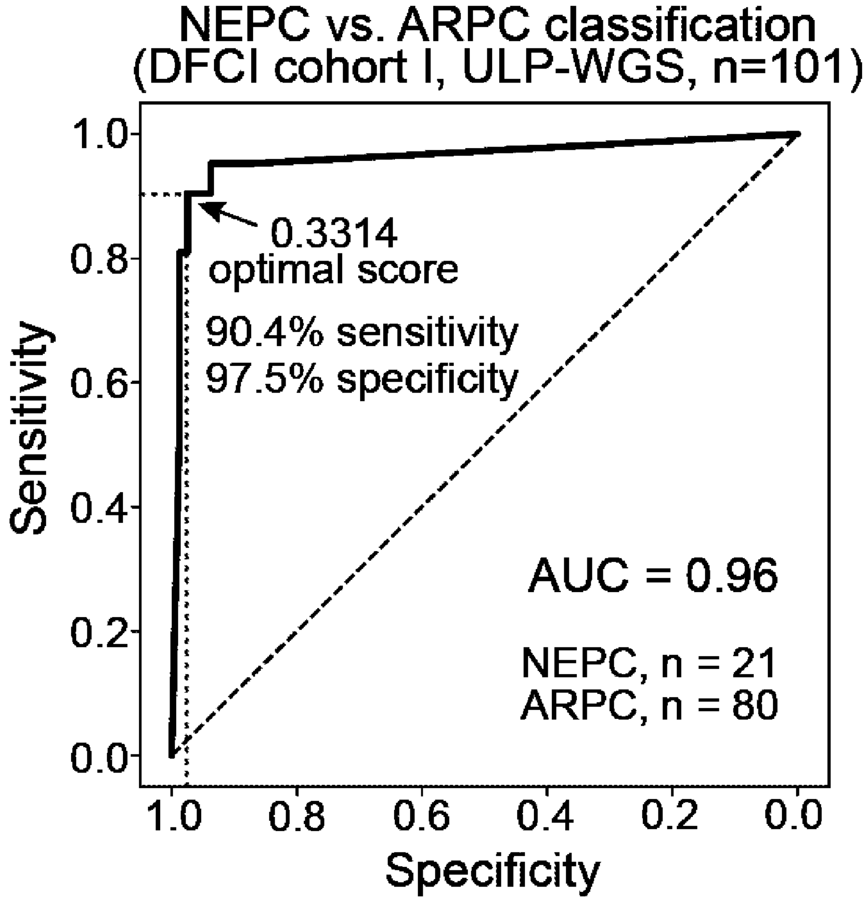

NEPC vs. ARPC classification
(DFCI cohort I, ULP-WGS, n=101)
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
0.3314
optimal score
90.4% sensitivity
97.5% specificity
AUC = 0.96
NEPC, n = 21
ARPC, n = 80

FIG. 15A

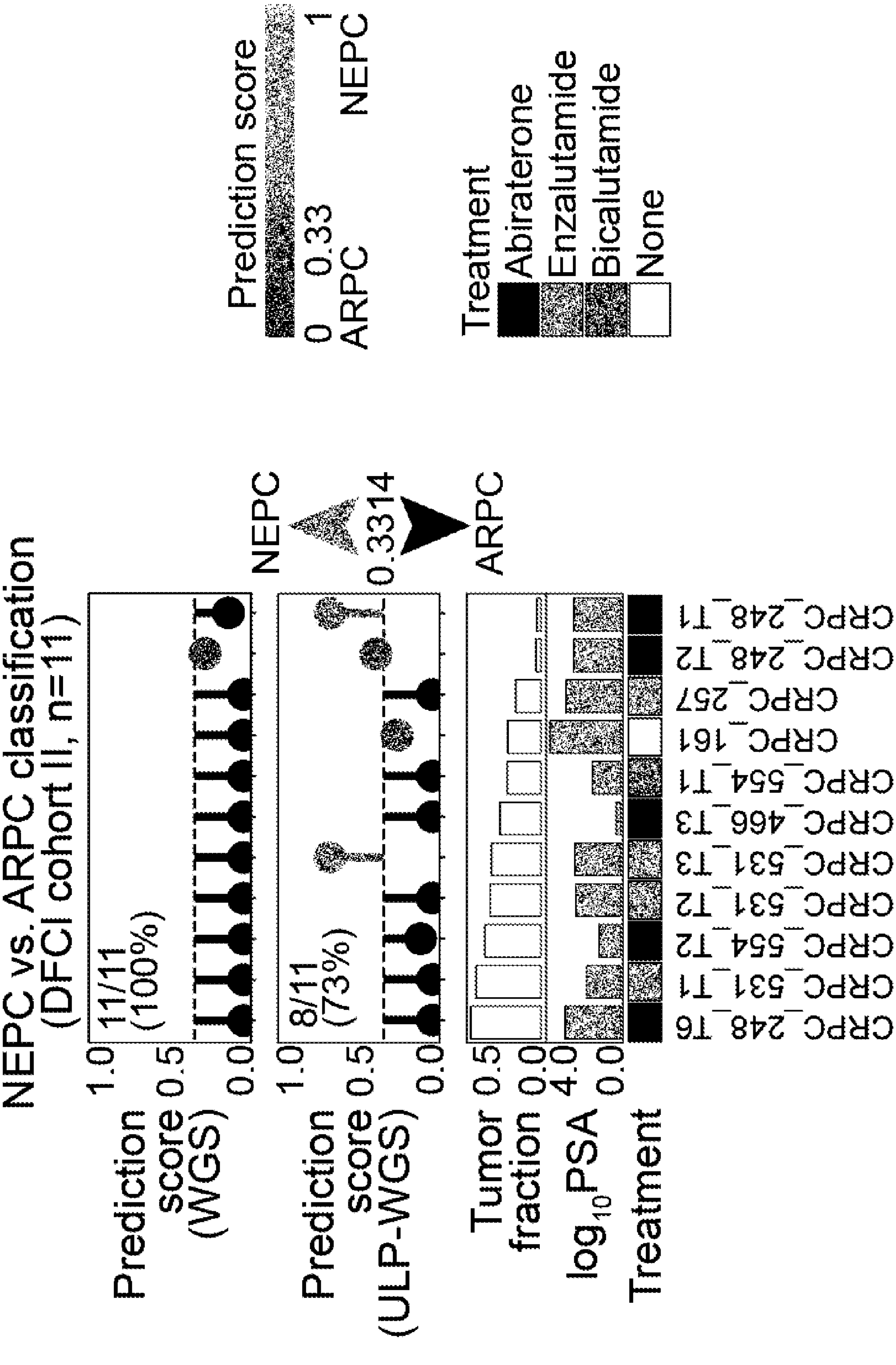

NEPC vs. ARPC classification
(DFCI cohort II, n=11)
Prediction score (WGS)
1.0
0.5
0.0
11/11 (100%)
Prediction score (ULP-WGS)
1.0
0.5
0.0
8/11 (73%)
NEPC
0.3314
ARPC
Tumor fraction
0.5
0.0
log10PSA
4.0
0.0
Treatment
CRPC_248_T6
CRPC_531_T1
CRPC_554_T2
CRPC_531_T2
CRPC_531_T3
CRPC_466_T3
CRPC_554_T1
CRPC_161
CRPC_257
CRPC_248_T2
CRPC_248_T1
Prediction score
0
0.33
1
ARPC
NEPC
Treatment
Abiraterone
Enzalutamide
Bicalutamide
None

FIG. 15B

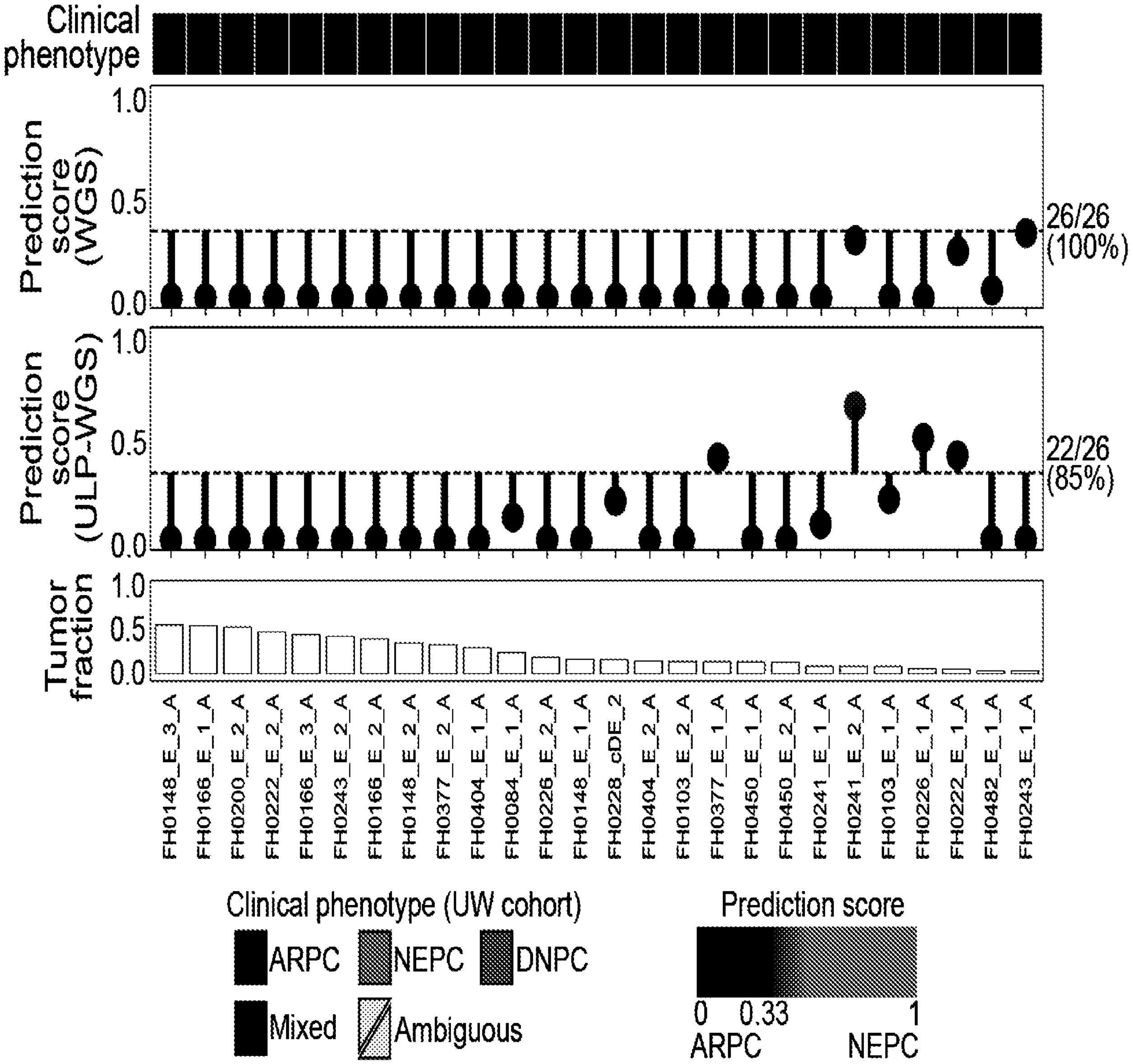

Clinical phenotype
Prediction score (WGS)
1.0
0.5
0.0
26/26
(100%)
Prediction score (ULP-WGS)
1.0
0.5
0.0
22/26
(85%)
Tumor fraction
1.0
0.5
0.0
FH0148_E_3_A
FH0166_E_1_A
FH0200_E_2_A
FH0222_E_2_A
FH0166_E_3_A
FH0243_E_2_A
FH0166_E_2_A
FH0148_E_2_A
FH0377_E_2_A
FH0404_E_1_A
FH0084_E_1_A
FH0226_E_2_A
FH0148_E_1_A
FH0228_cDE_2
FH0404_E_2_A
FH0103_E_2_A
FH0377_E_1_A
FH0450_E_1_A
FH0450_E_2_A
FH0241_E_1_A
FH0241_E_2_A
FH0103_E_1_A
FH0226_E_1_A
FH0222_E_1_A
FH0482_E_1_A
FH0243_E_1_A
Clinical phenotype (UW cohort)
ARPC
NEPC
DNPC
Mixed
Ambiguous
Prediction score
0
0.33
1
ARPC
NEPC

FIG. 15C

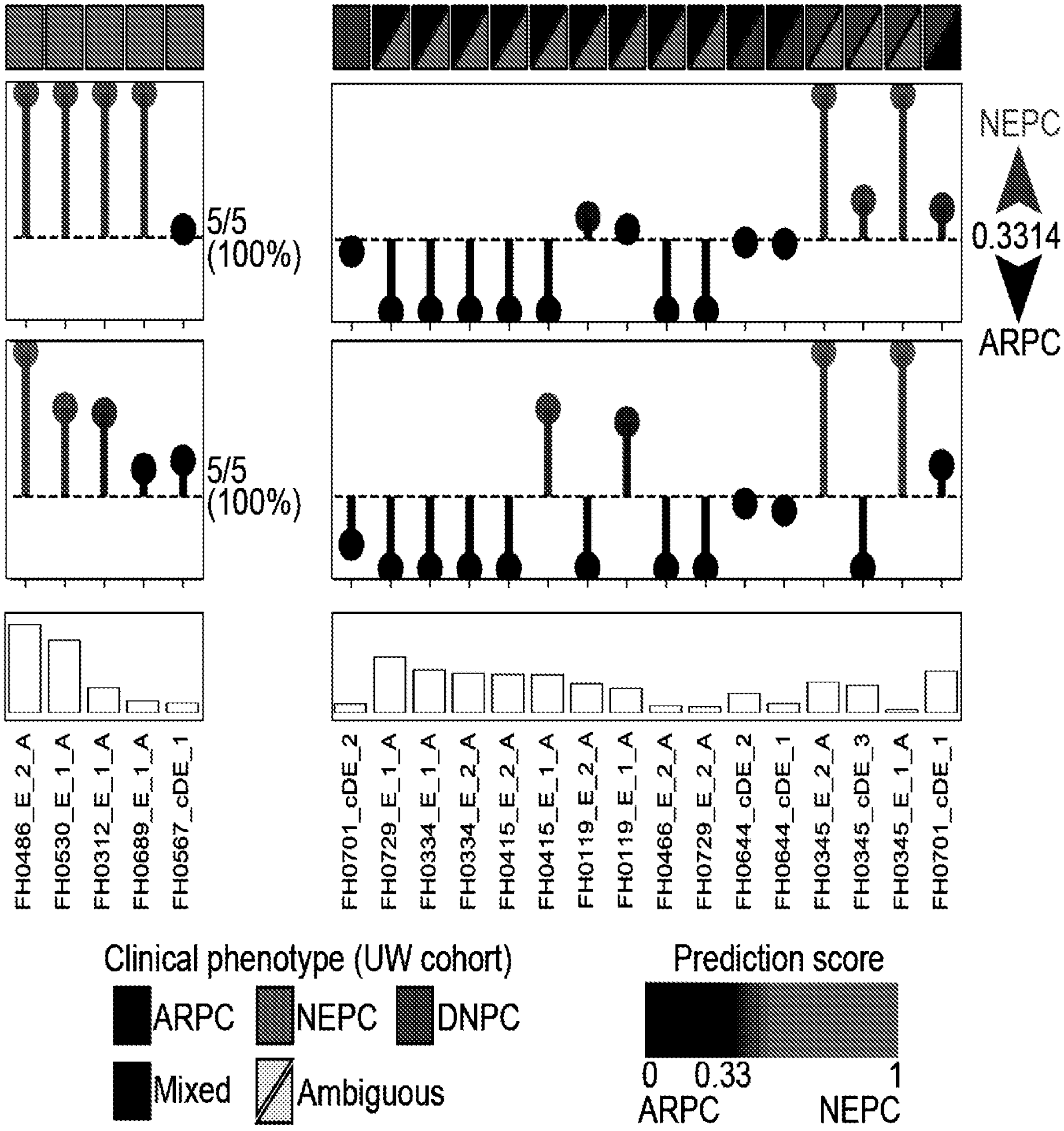

NEPC
0.3314
ARPC
5/5 (100%)
5/5 (100%)
FH0486_E_2_A
FH0530_E_1_A
FH0312_E_1_A
FH0689_E_1_A
FH0567_cDE_1
FH0701_cDE_2
FH0729_E_1_A
FH0334_E_1_A
FH0334_E_2_A
FH0415_E_2_A
FH0415_E_1_A
FH0119_E_2_A
FH0119_E_1_A
FH0466_E_2_A
FH0729_E_2_A
FH0644_cDE_2
FH0644_cDE_1
FH0345_E_2_A
FH0345_cDE_3
FH0345_E_1_A
FH0701_cDE_1
Clinical phenotype (UW cohort)
ARPC
NEPC
DNPC
Mixed
Ambiguous
Prediction score
0
0.33
1
ARPC
NEPC

FIG. 15C
(CONT.)

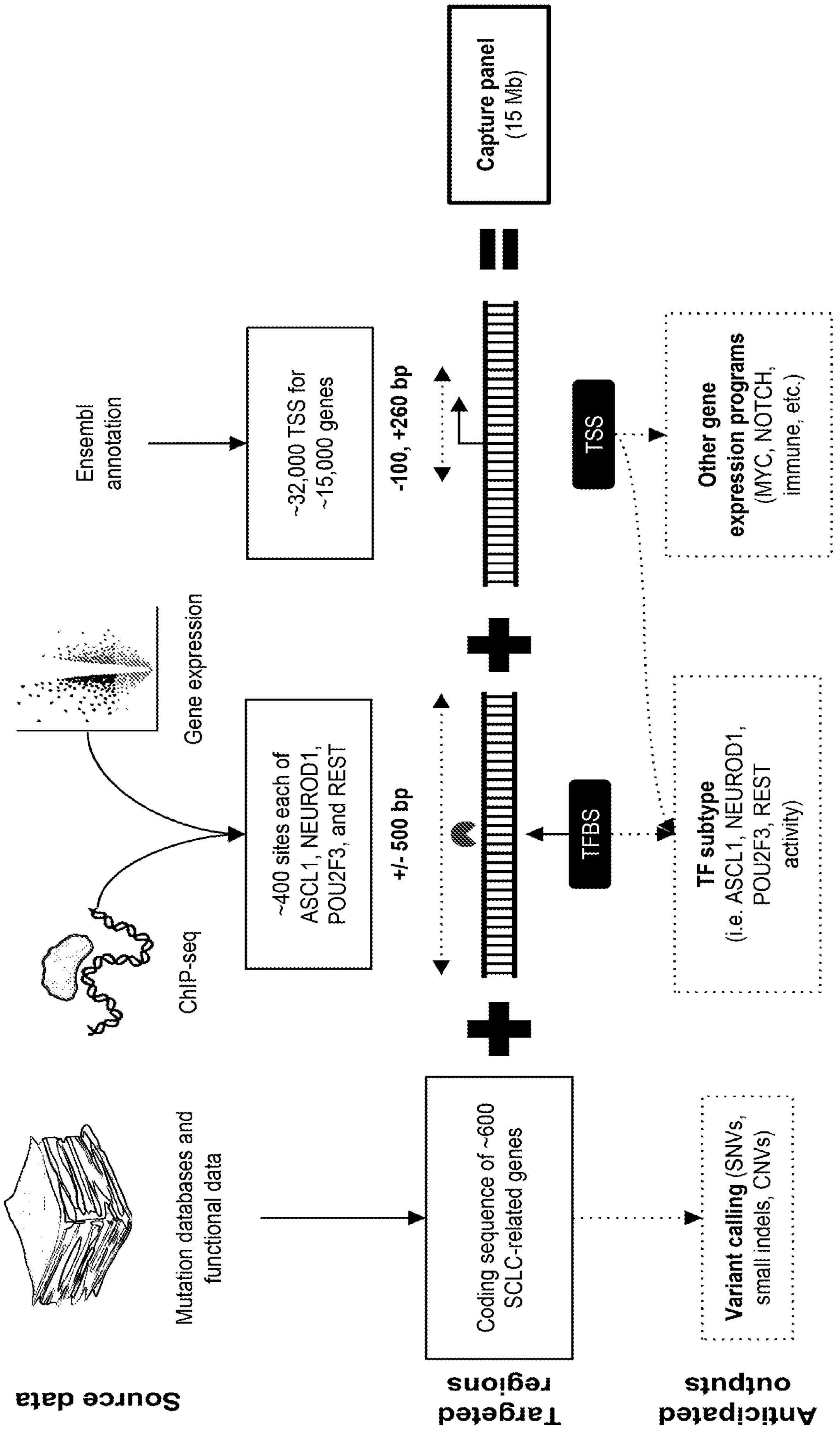

Source data
Mutation databases and functional data
ChIP-seq
Gene expression
Ensembl annotation
Targeted regions
Coding sequence of ~600 SCLC-related genes
~400 sites each of ASCL1, NEUROD1, POU2F3, and REST
+/- 500 bp
TFBS
~32,000 TSS for ~15,000 genes
-100, +260 bp
TSS
Capture panel
(15 Mb)
Anticipated outputs
Variant calling (SNVs, small indels, CNVs)
TF subtype
(i.e. ASCL1, NEUROD1, POU2F3, REST activity)
Other gene expression programs
(MYC, NOTCH, immune, etc.)

FIG. 16

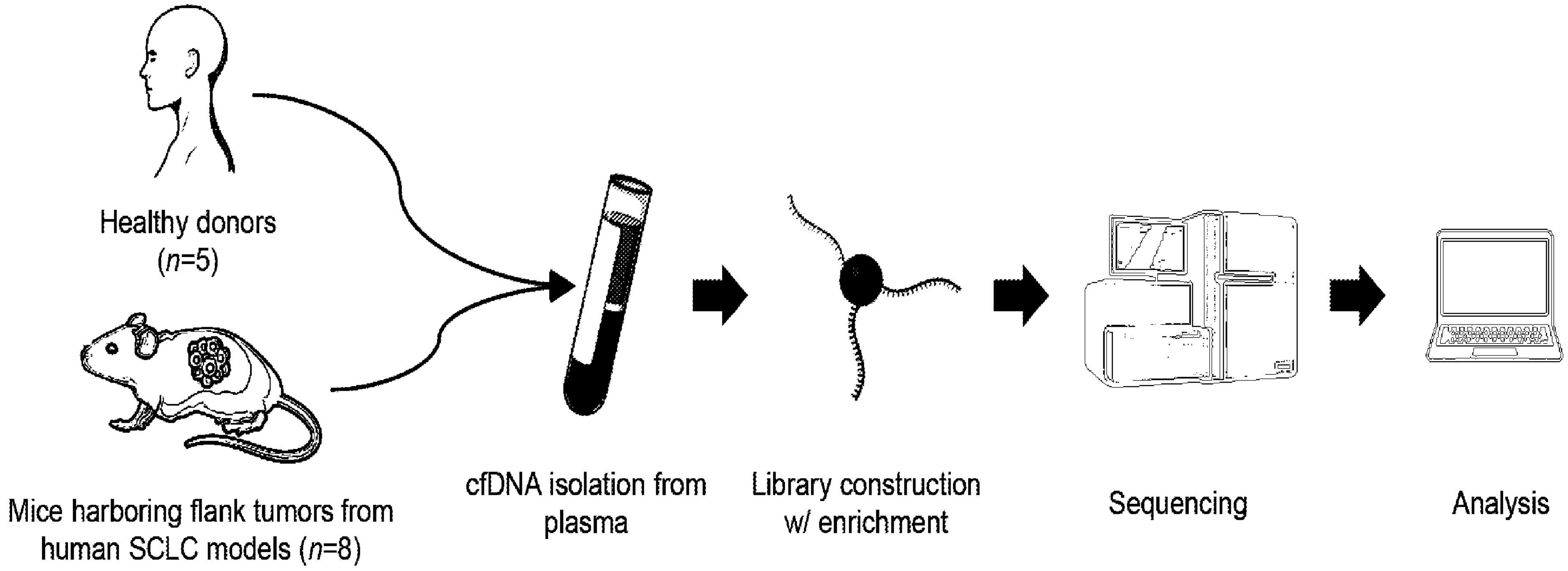

Healthy donors
(*n*=5)
Mice harboring flank tumors from
human SCLC models (*n*=8)
cfDNA isolation from
plasma
Library construction
w/ enrichment
Sequencing
Analysis

FIG. 17A

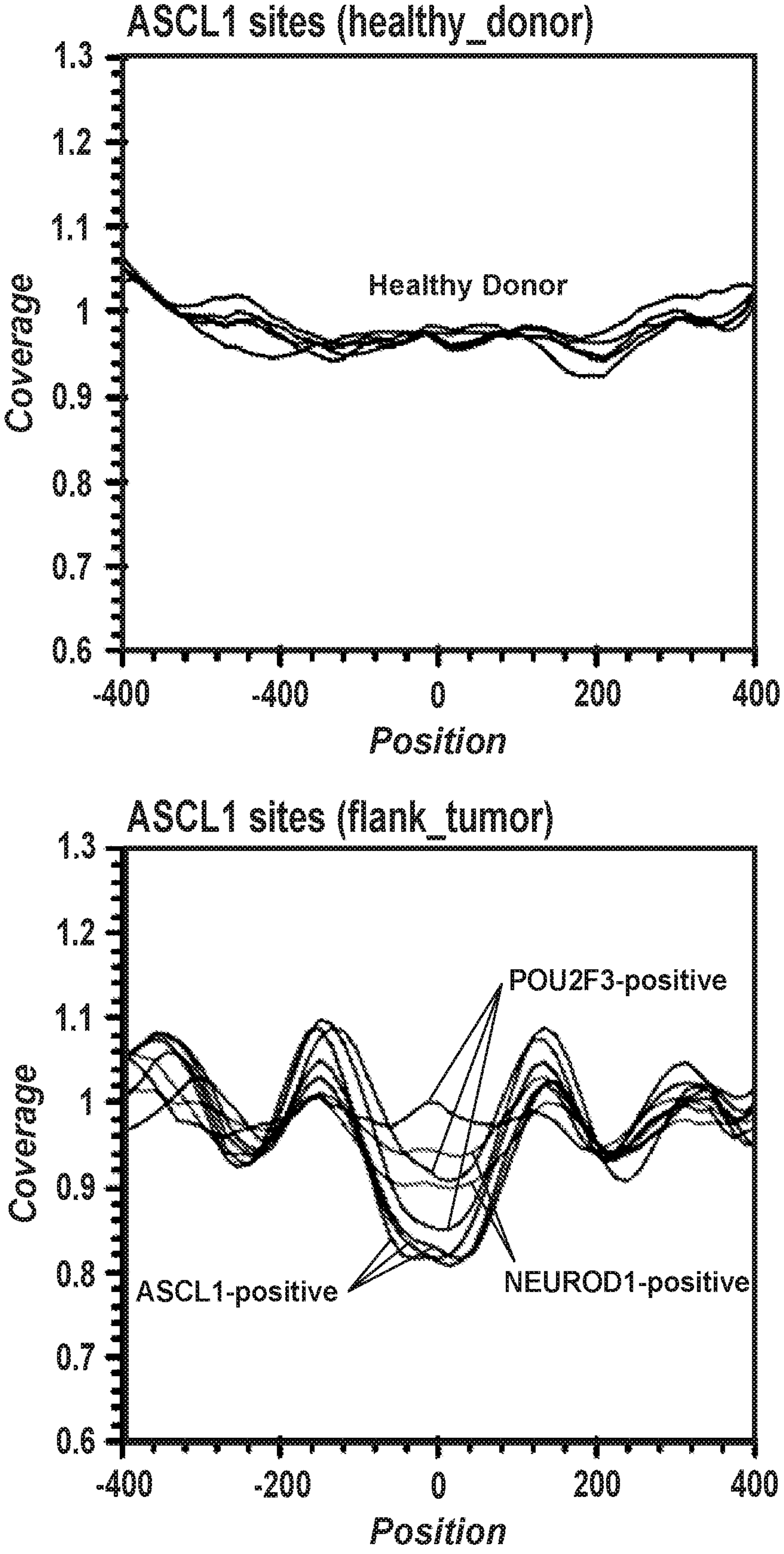
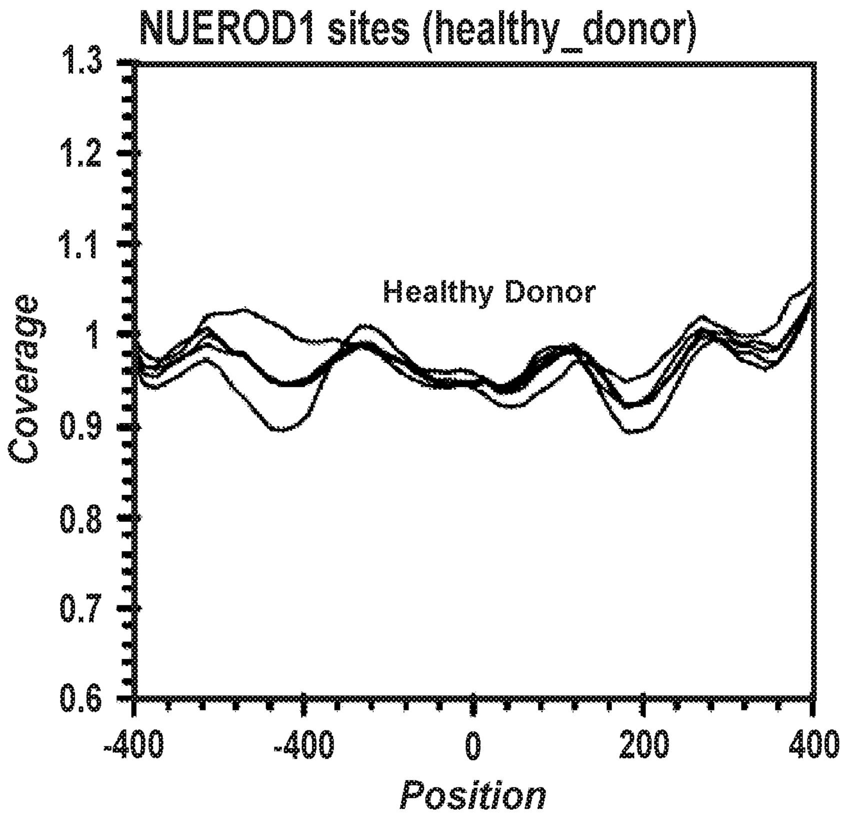

ASCL1 sites (healthy_donor)
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
Coverage
Healthy Donor
-400
-400
0
200
400
Position
ASCL1 sites (flank_tumor)
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
Coverage
POU2F3-positive
ASCL1-positive
NEUROD1-positive
-400
-200
0
200
400
Position

*FIG. 17B*


NUEROD1 sites (healthy_donor)
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
Coverage
Healthy Donor
-400
-400
0
200
400
Position

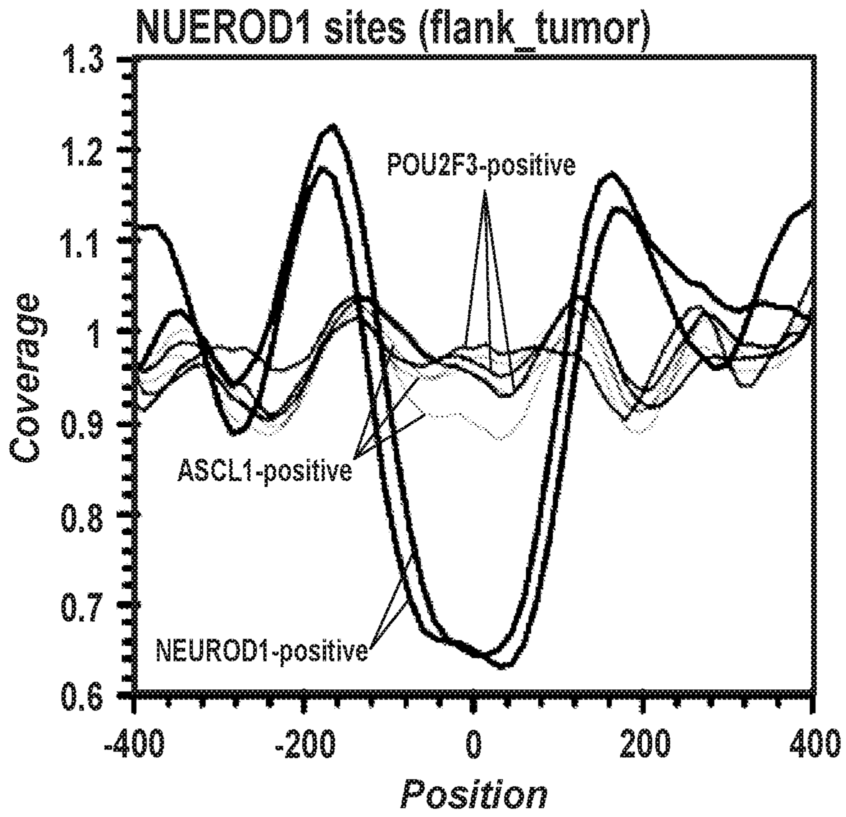
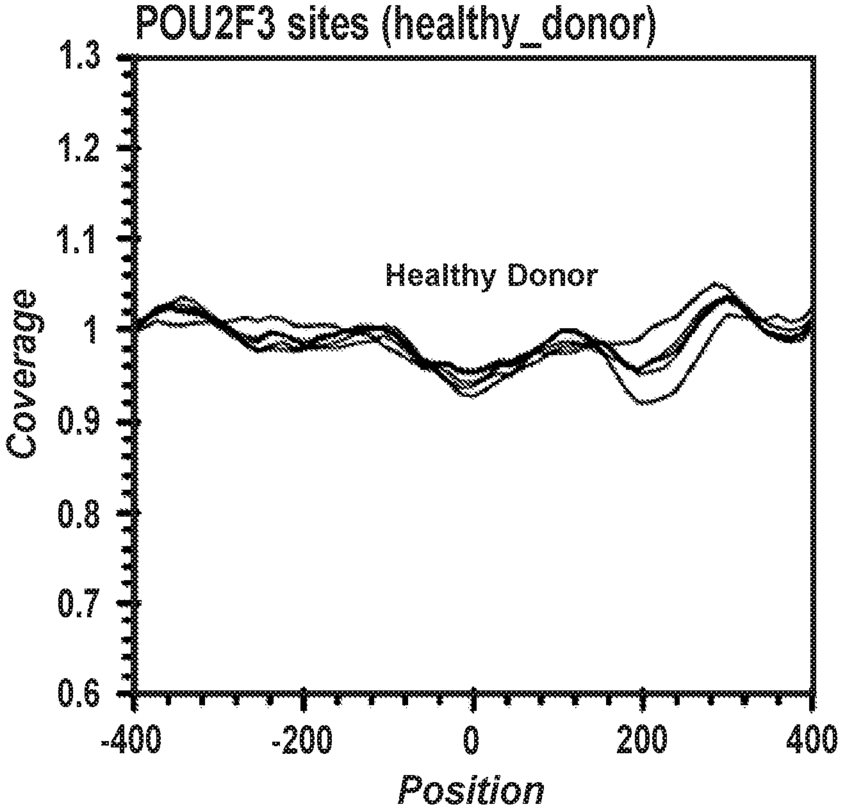

NUEROD1 sites (flank_tumor)
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
Coverage
POU2F3-positive
ASCL1-positive
NEUROD1-positive
-400
-200
0
200
400
Position

*FIG. 17B*
*(CONT.)*


POU2F3 sites (healthy_donor)
Healthy Donor
Coverage
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
-400
-200
0
200
400
Position

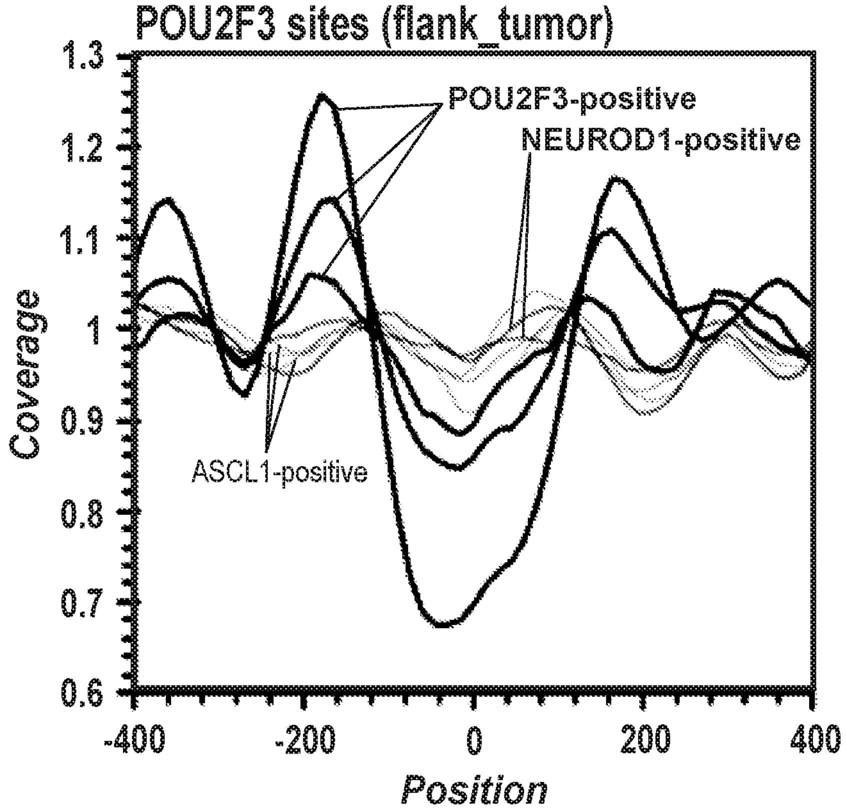
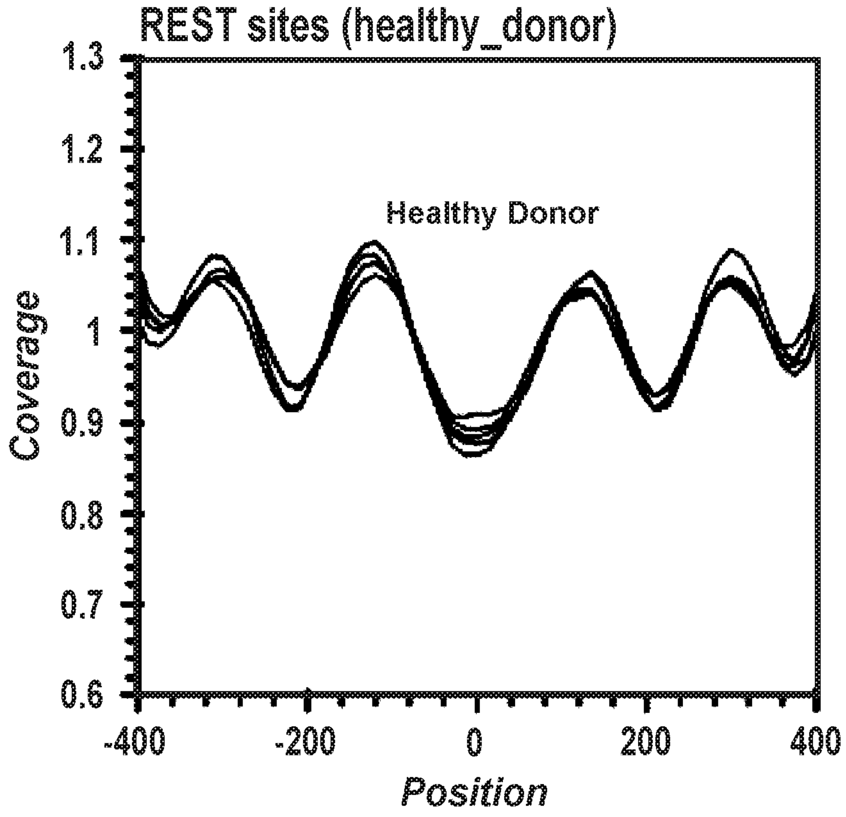

POU2F3 sites (flank_tumor)
POU2F3-positive
NEUROD1-positive
ASCL1-positive
Coverage
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
-400
-200
0
200
400
Position

FIG. 17B

(CONT.)


REST sites (healthy_donor)
Healthy Donor
Coverage
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
-400
-200
0
200
400
Position

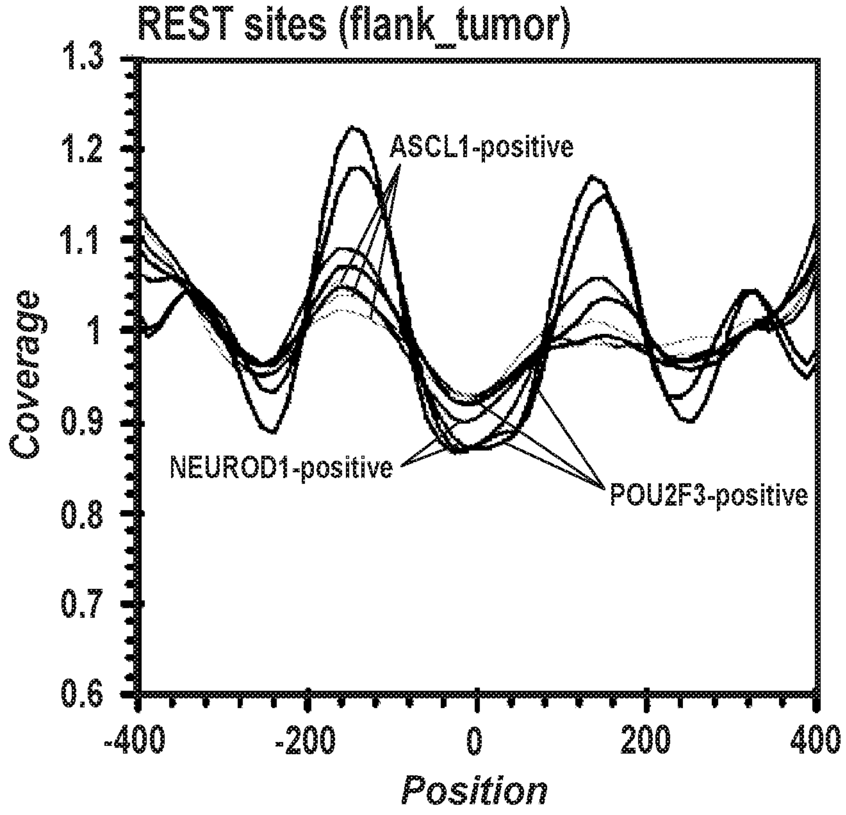

REST sites (flank_tumor)
ASCL1-positive
NEUROD1-positive
POU2F3-positive
Coverage
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
-400
-200
0
200
400
Position

FIG. 17B
(CONT.)

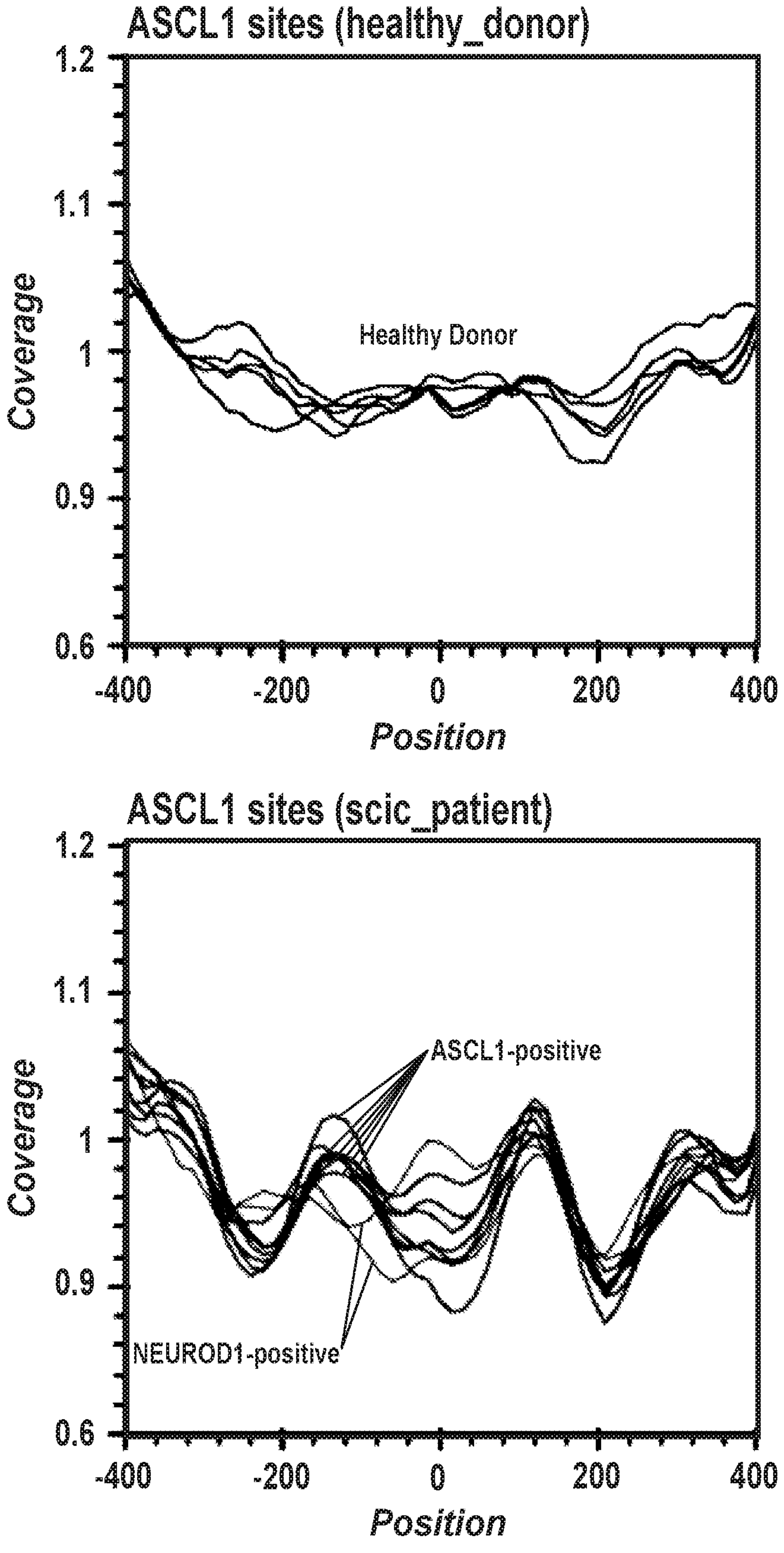

ASCL1 sites (healthy_donor)
1.2
1.1
1
0.9
0.6
Coverage
Healthy Donor
-400
-200
0
200
400
Position
ASCL1 sites (scic_patient)
1.2
1.1
1
0.9
0.6
Coverage
ASCL1-positive
NEUROD1-positive
-400
-200
0
200
400
Position

FIG. 18A

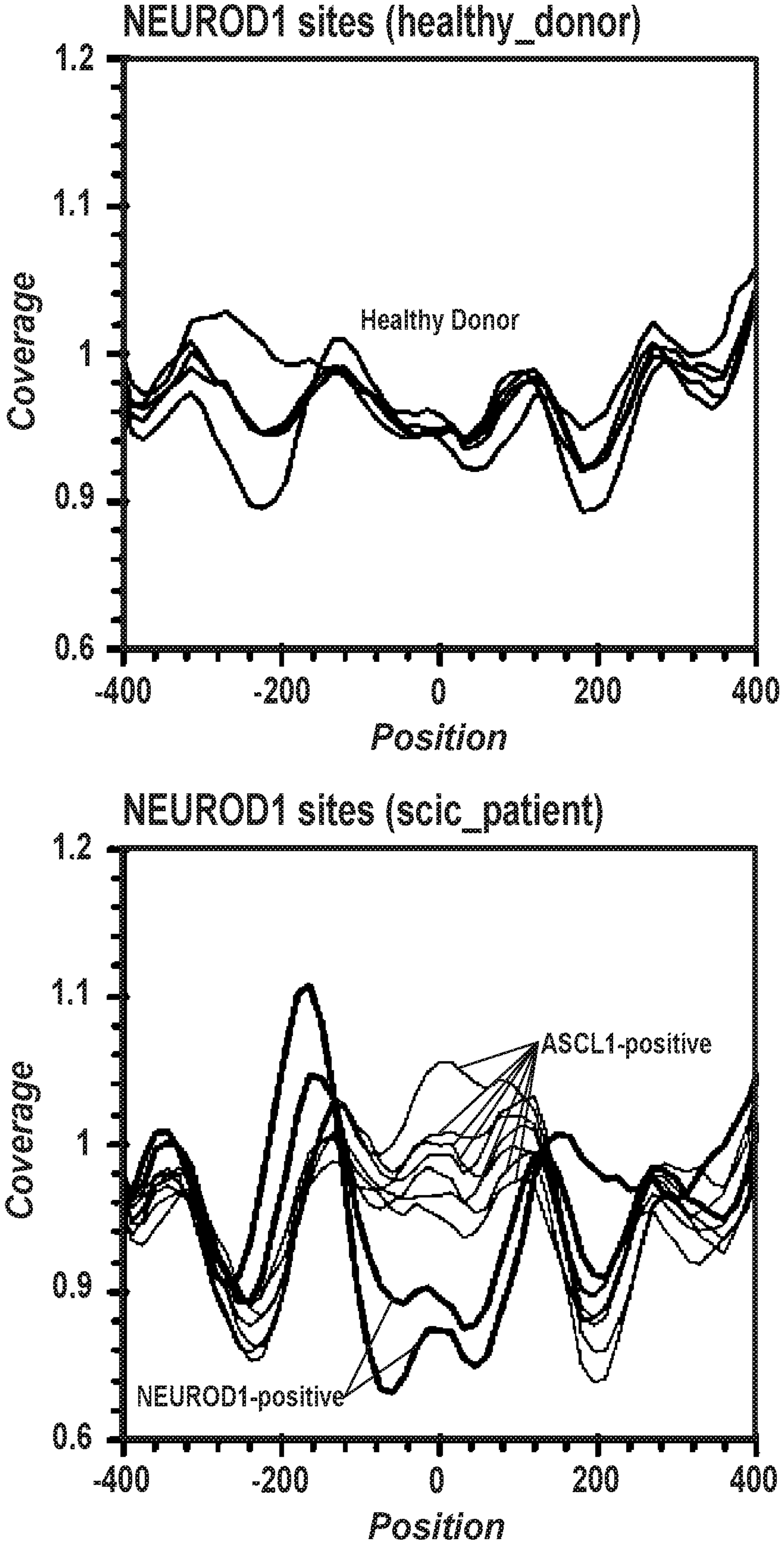
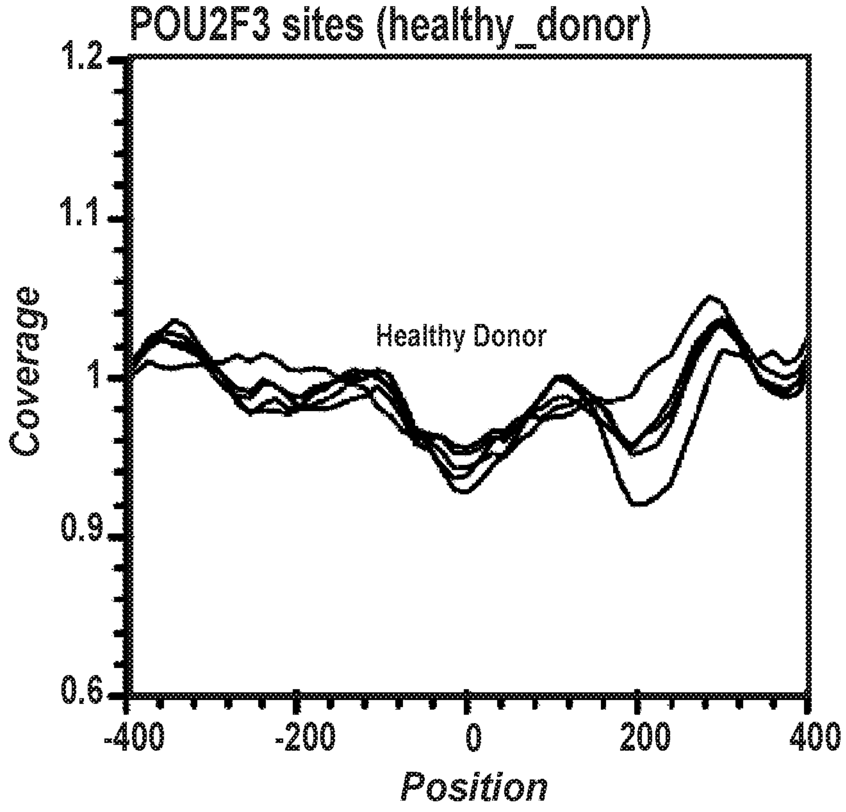

NEUROD1 sites (healthy_donor)
Coverage
1.2
1.1
1
0.9
0.6
-400
-200
0
200
400
Position
Healthy Donor
NEUROD1 sites (scic_patient)
Coverage
1.2
1.1
1
0.9
0.6
-400
-200
0
200
400
Position
ASCL1-positive
NEUROD1-positive

FIG. 18A
(CONT.)


POU2F3 sites (healthy_donor)
Coverage
Position
1.2
1.1
1
0.9
0.6
-400
-200
0
200
400
Healthy Donor

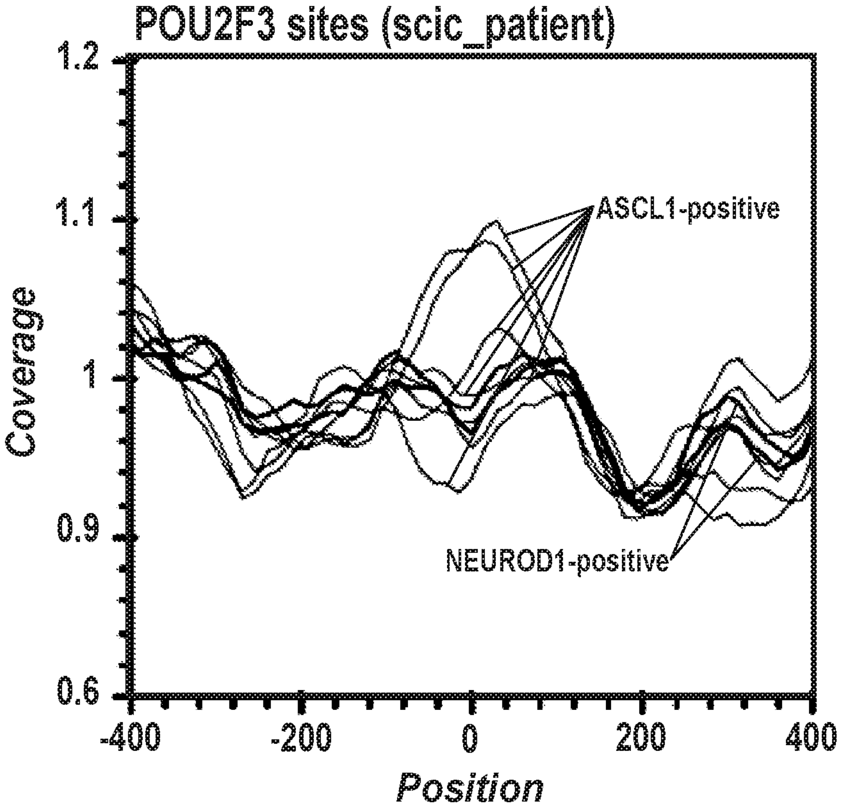

POU2F3 sites (scic_patient)
Coverage
Position
1.2
1.1
1
0.9
0.6
-400
-200
0
200
400
ASCL1-positive
NEUROD1-positive

FIG. 18A
(CONT.)

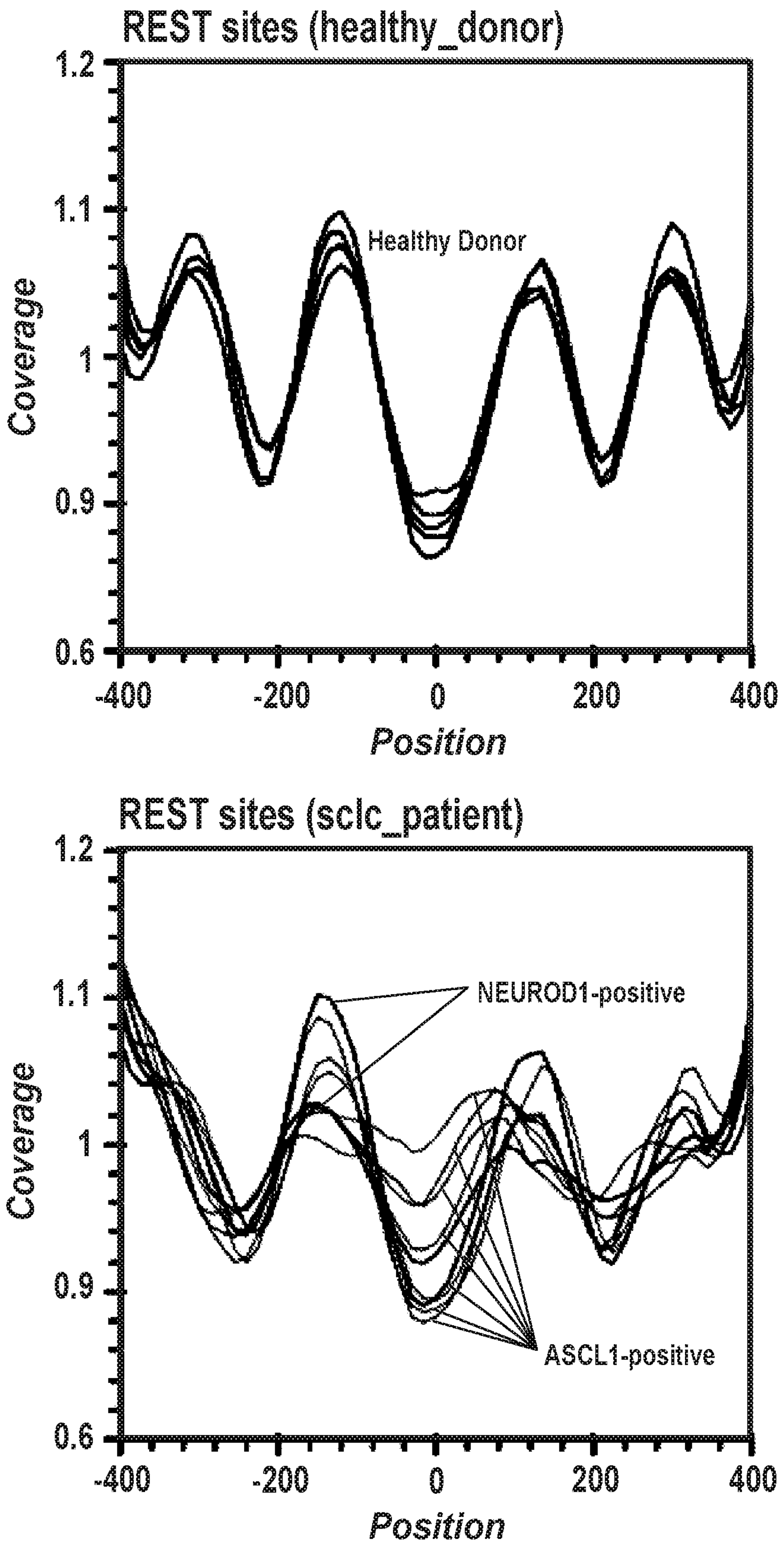

REST sites (healthy_donor)
Healthy Donor
Coverage
1.2
1.1
1
0.9
0.6
-400
-200
0
200
400
Position
REST sites (sclc_patient)
NEUROD1-positive
ASCL1-positive
Coverage
1.2
1.1
1
0.9
0.6
-400
-200
0
200
400
Position

*FIG. 18A*
*(CONT.)*

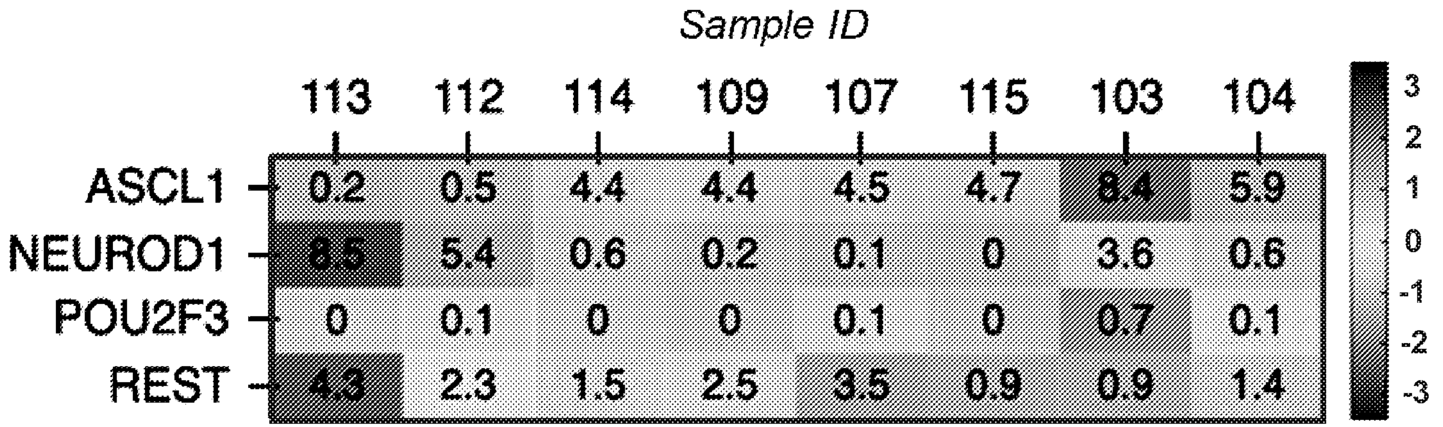


FIG. 18B

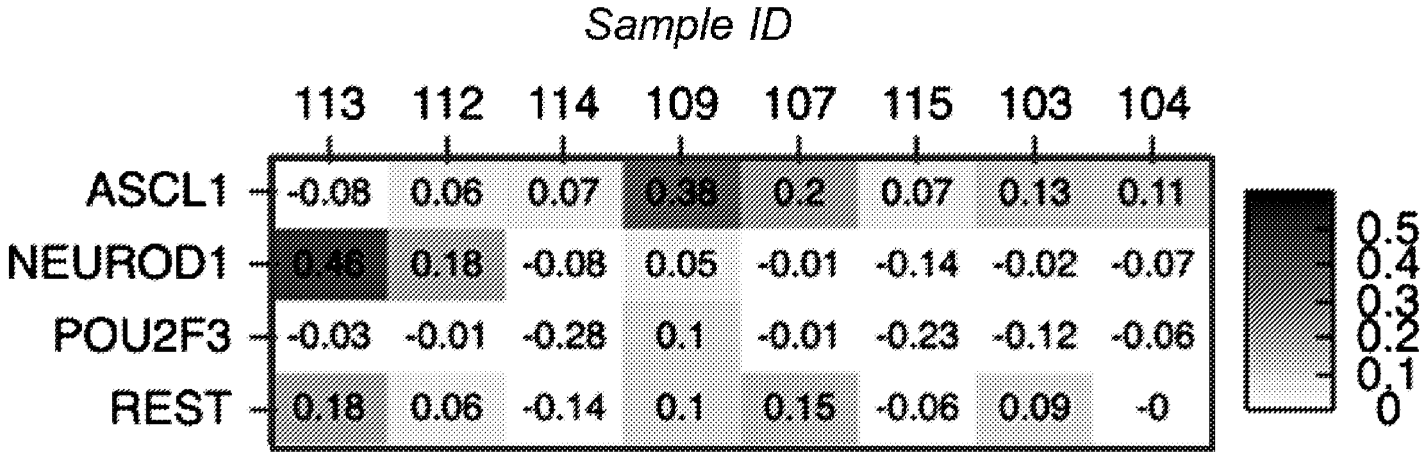


FIG. 18C

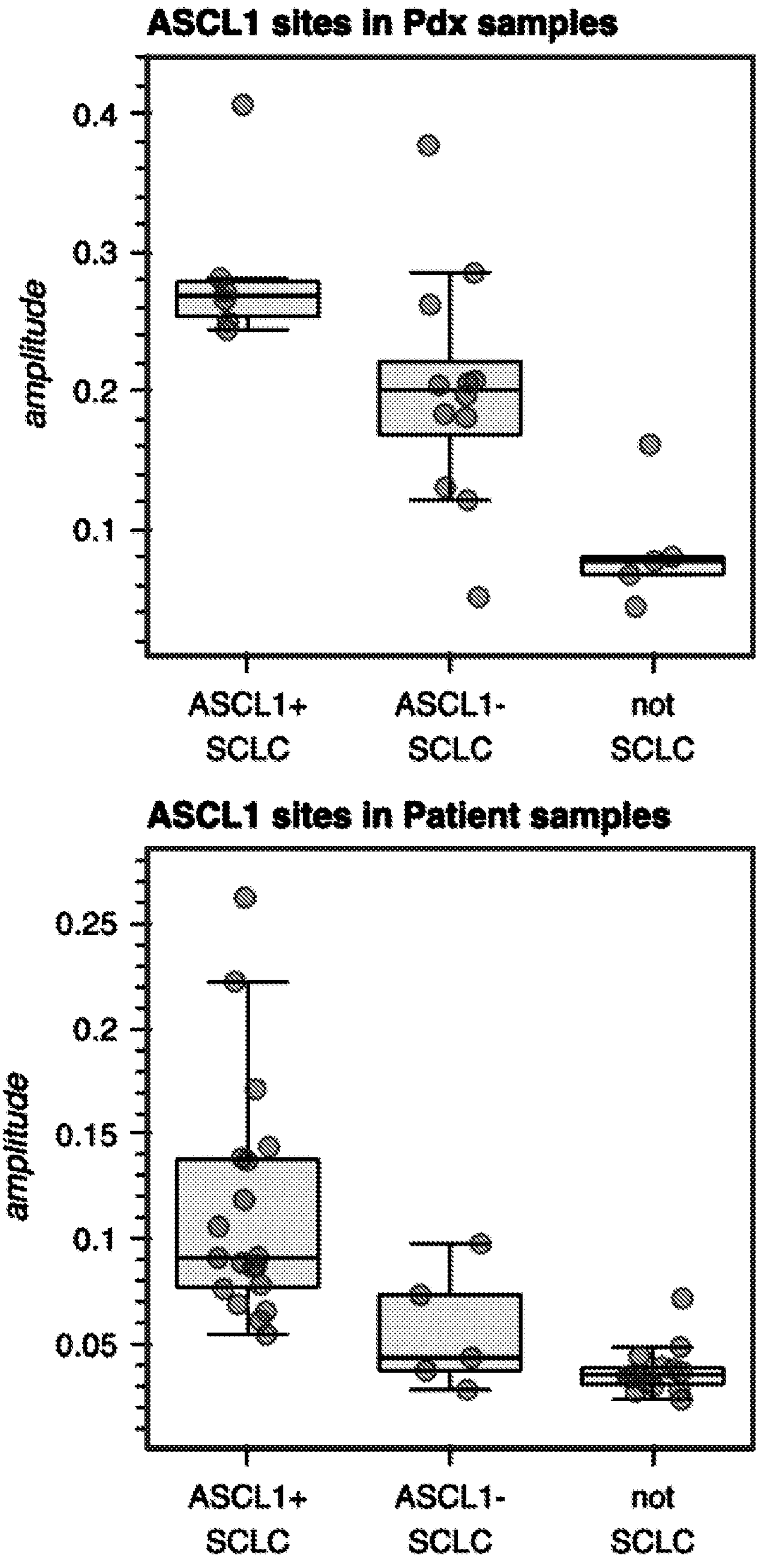
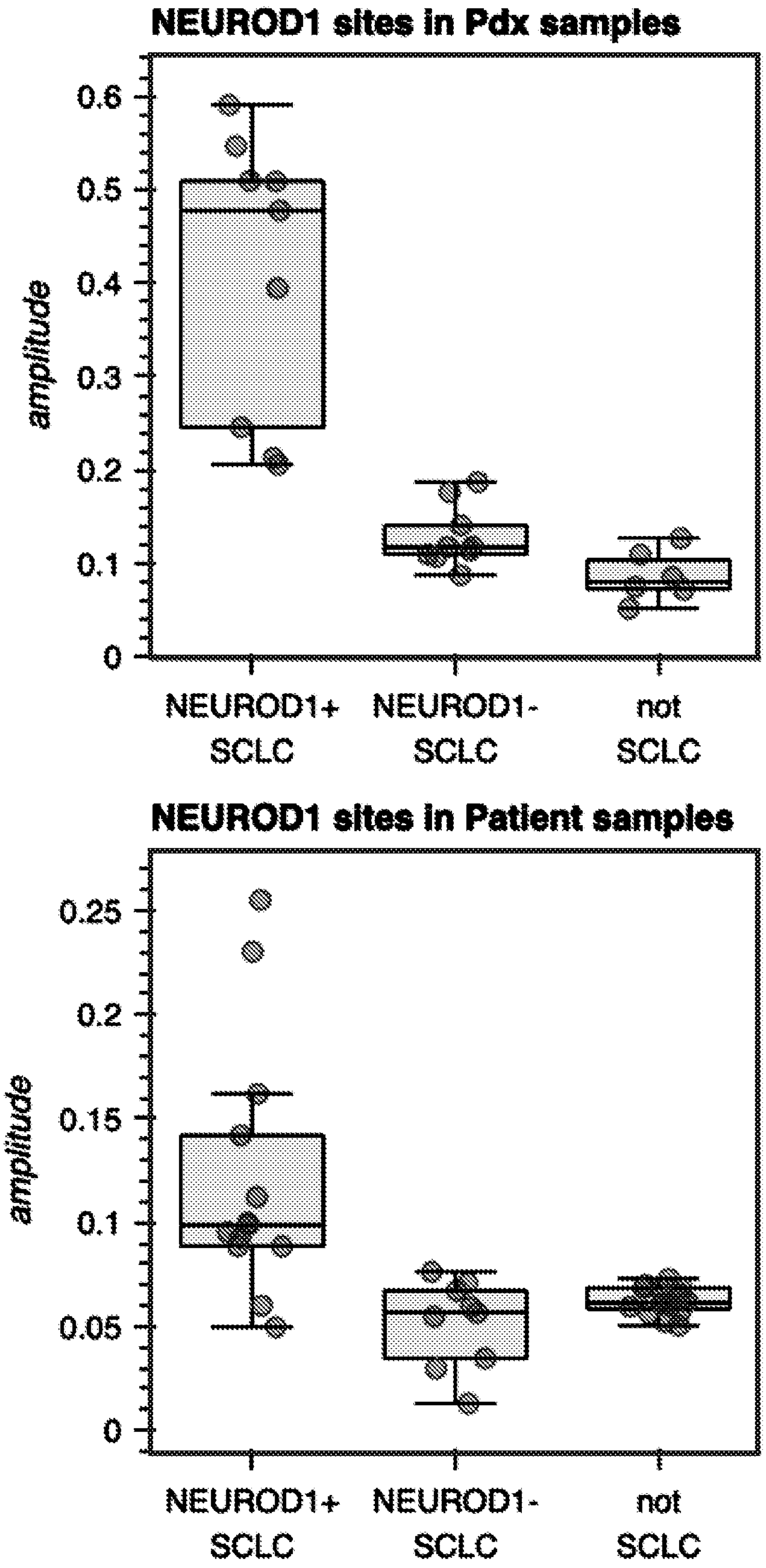

ASCL1 sites in Pdx samples
amplitude
0.4
0.3
0.2
0.1
ASCL1+ SCLC
ASCL1- SCLC
not SCLC
ASCL1 sites in Patient samples
amplitude
0.25
0.2
0.15
0.1
0.05
ASCL1+ SCLC
ASCL1- SCLC
not SCLC

*FIG. 19*

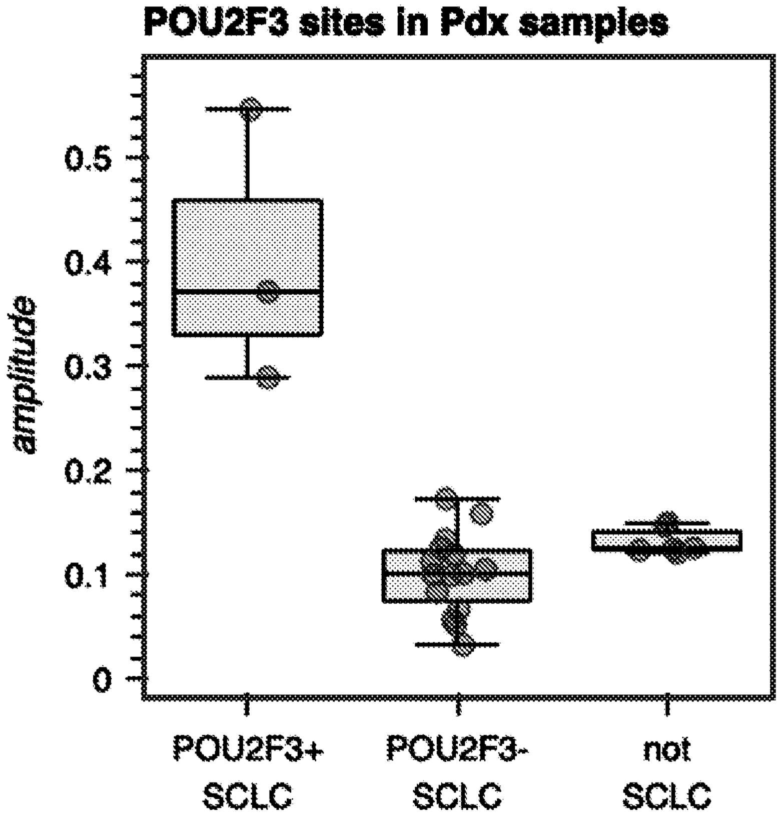


NEUROD1 sites in Pdx samples

NEUROD1 sites in Patient samples

*FIG. 19*

*(CONT.)*


POU2F3 sites in Pdx samples
amplitude
0.5
0.4
0.3
0.2
0.1
0
POU2F3+
SCLC
POU2F3-
SCLC
not
SCLC

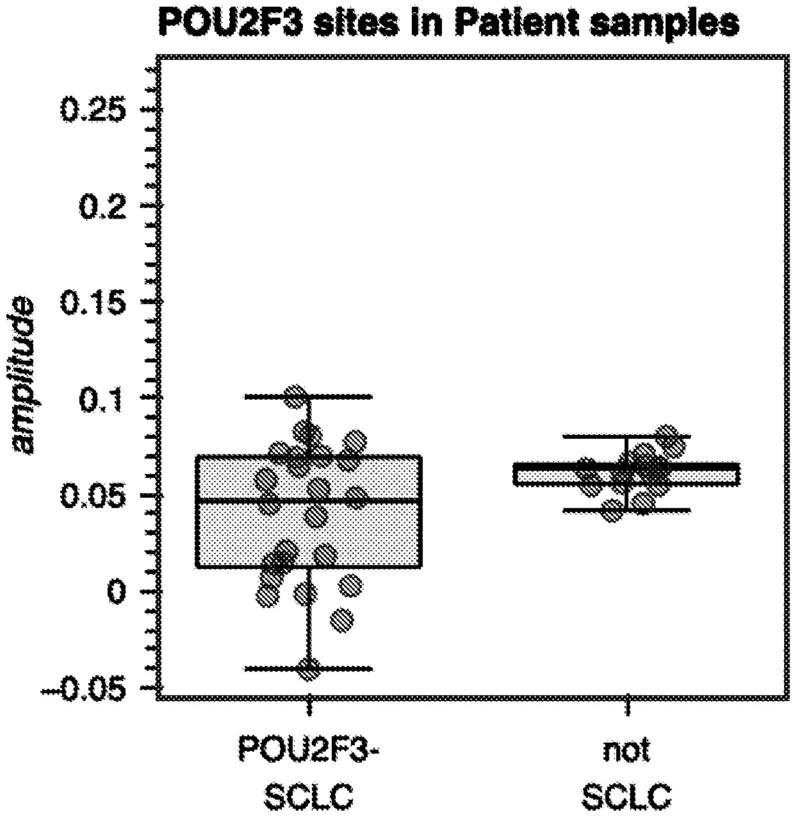

POU2F3 sites in Patient samples
amplitude
0.25
0.2
0.15
0.1
0.05
0
−0.05
POU2F3-
SCLC
not
SCLC

FIG. 19
(CONT.)

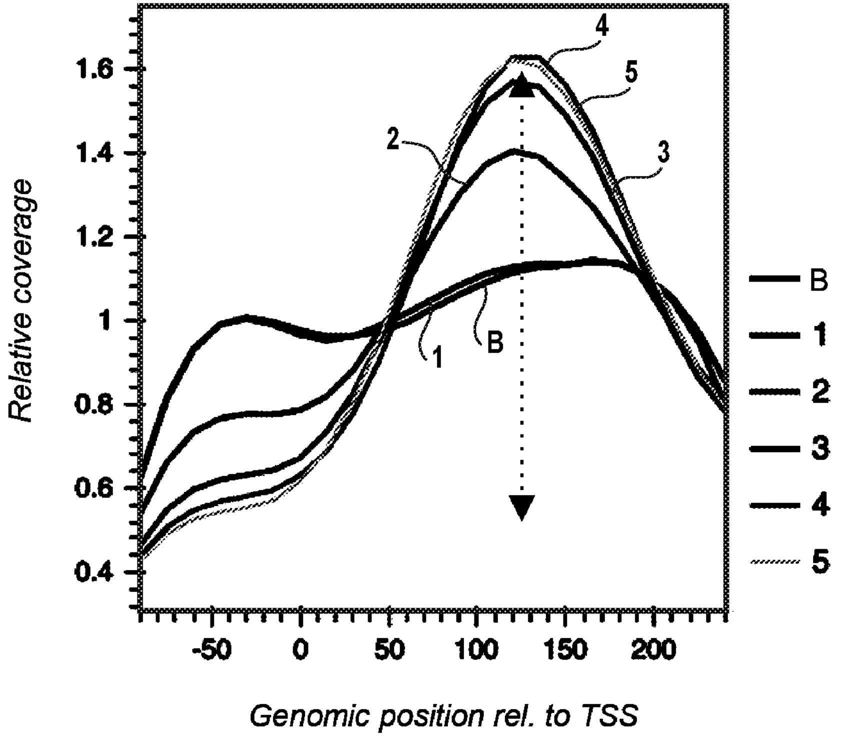


FIG. 20A

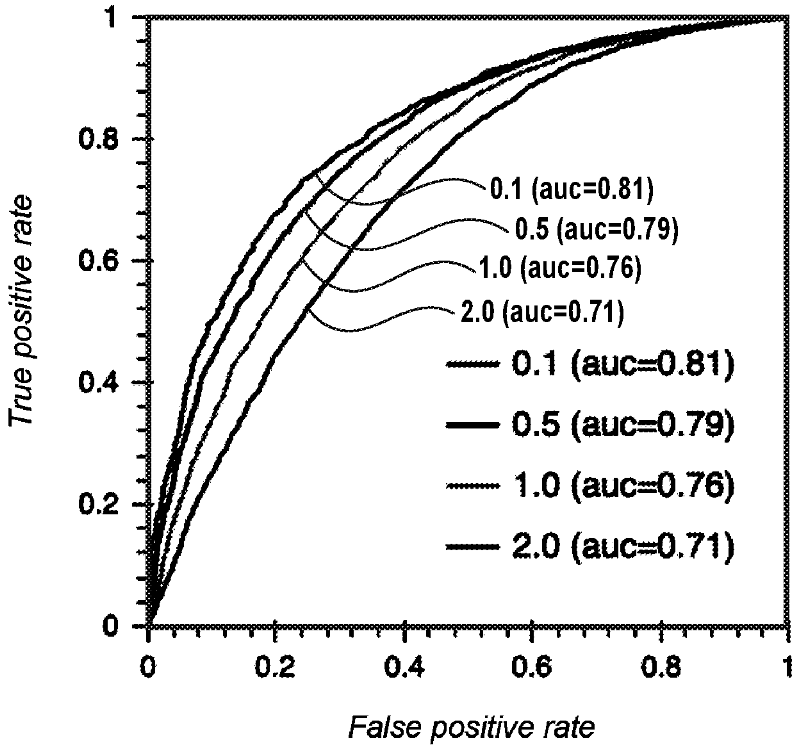


FIG. 20B

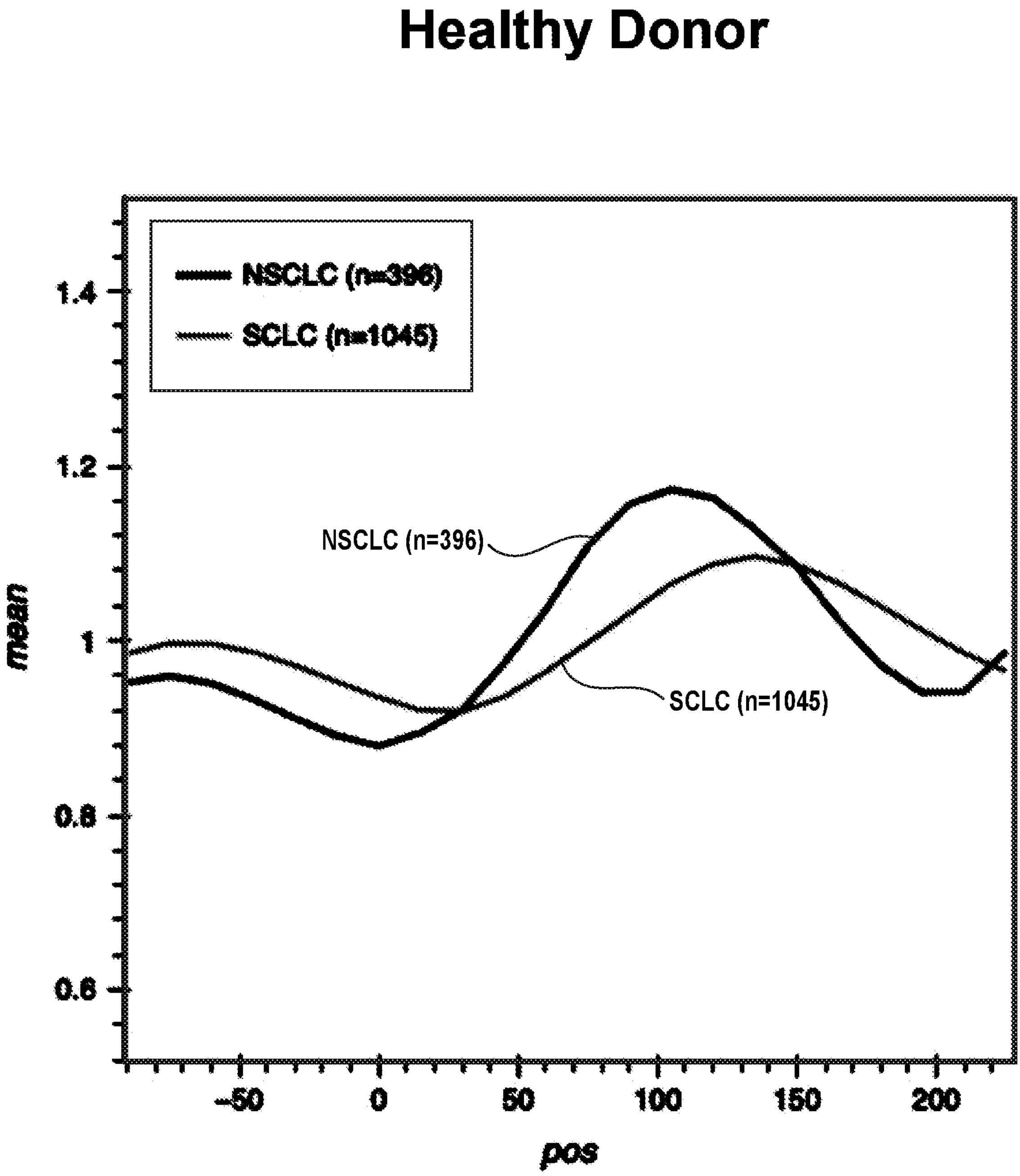

Healthy Donor
NSCLC (n=396)
SCLC (n=1045)
1.4
1.2
1
0.8
0.6
mean
-50
0
50
100
150
200
pos
NSCLC (n=396)
SCLC (n=1045)

FIG. 21A

**NSCLC PDX** (MSK_LX631)

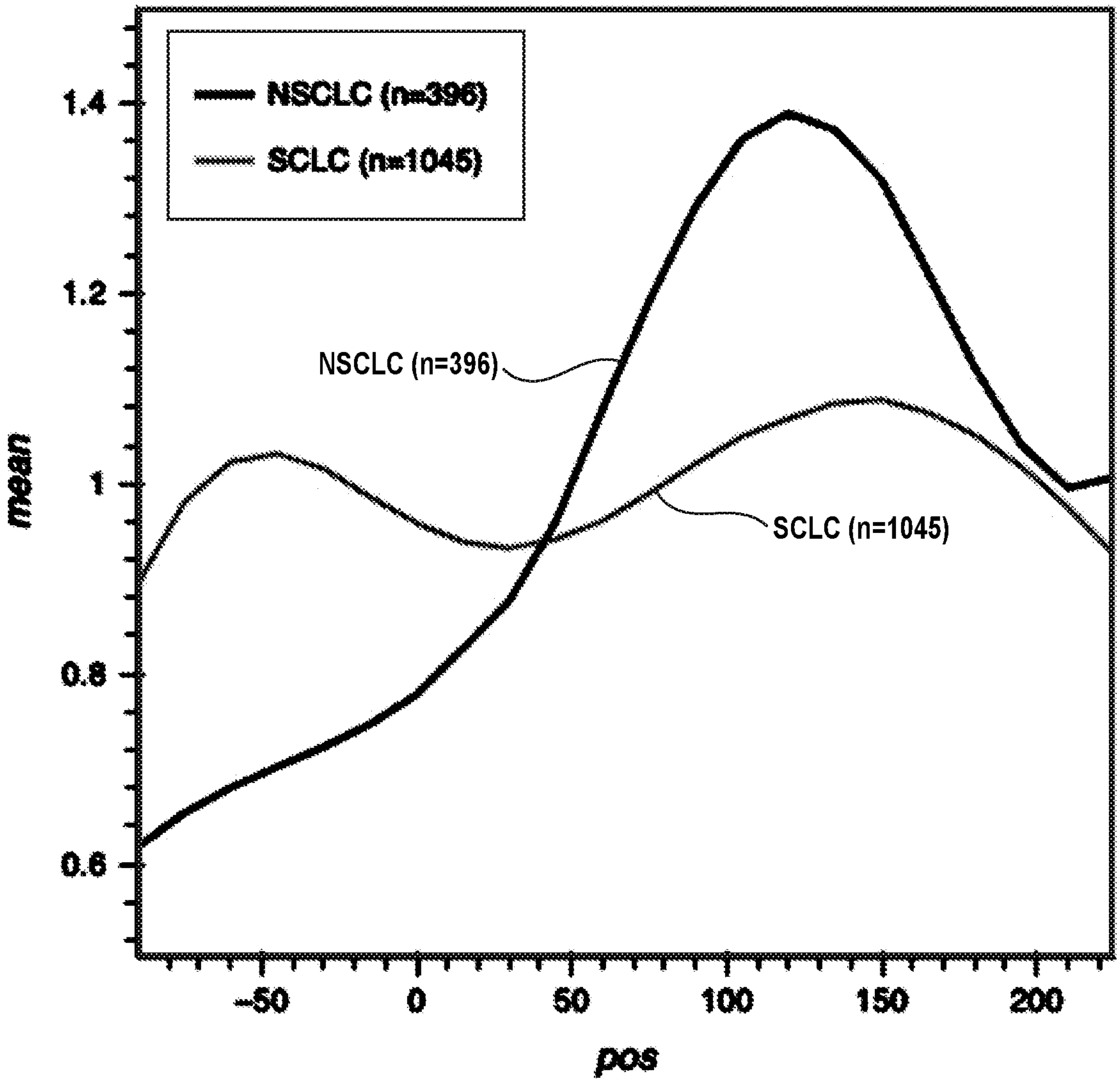


*FIG. 21B*

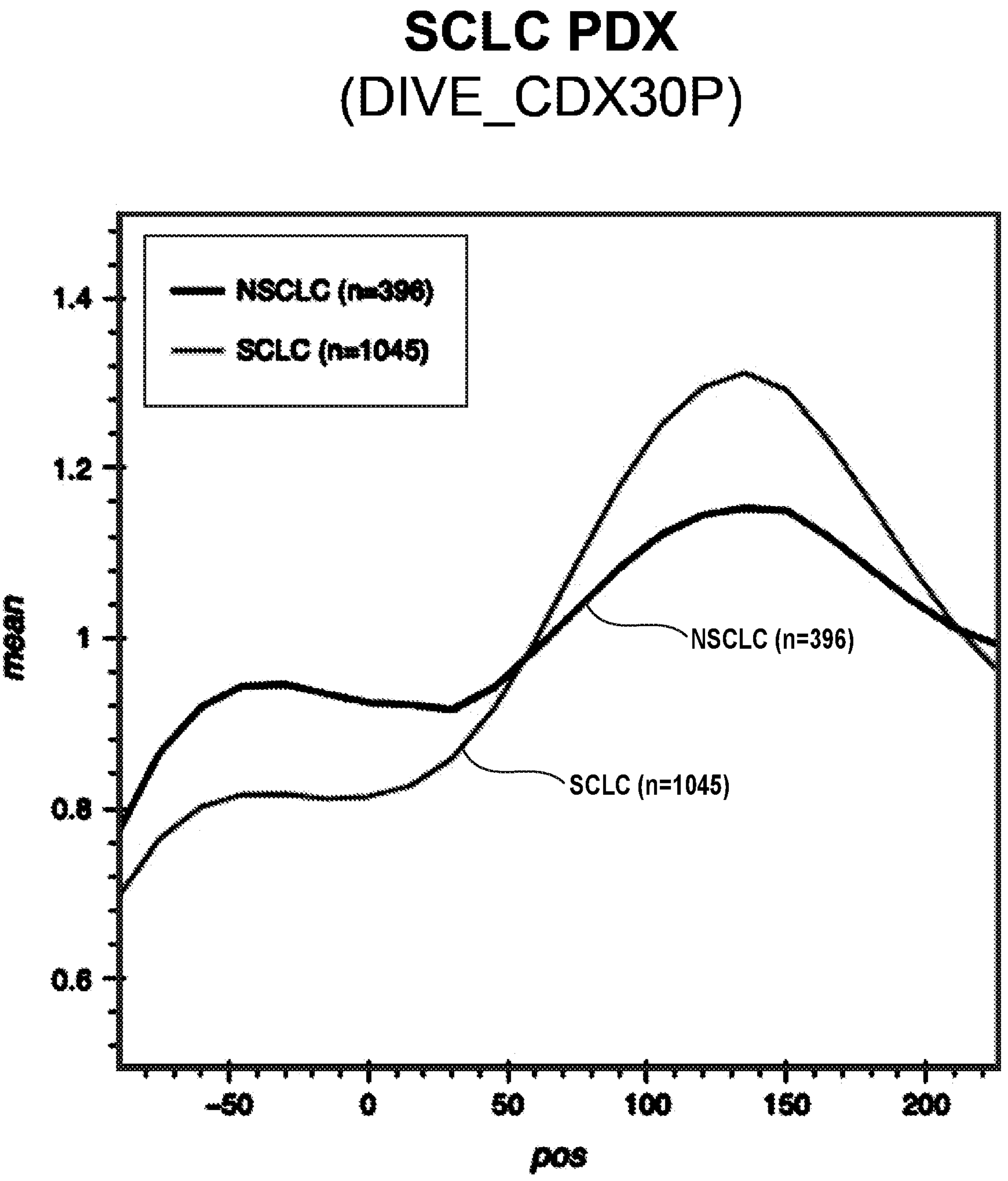

SCLC PDX
(DIVE_CDX30P)
NSCLC (n=396)
SCLC (n=1045)
1.4
1.2
1
0.8
0.6
mean
-50
0
50
100
150
200
pos
NSCLC (n=396)
SCLC (n=1045)

FIG. 21C

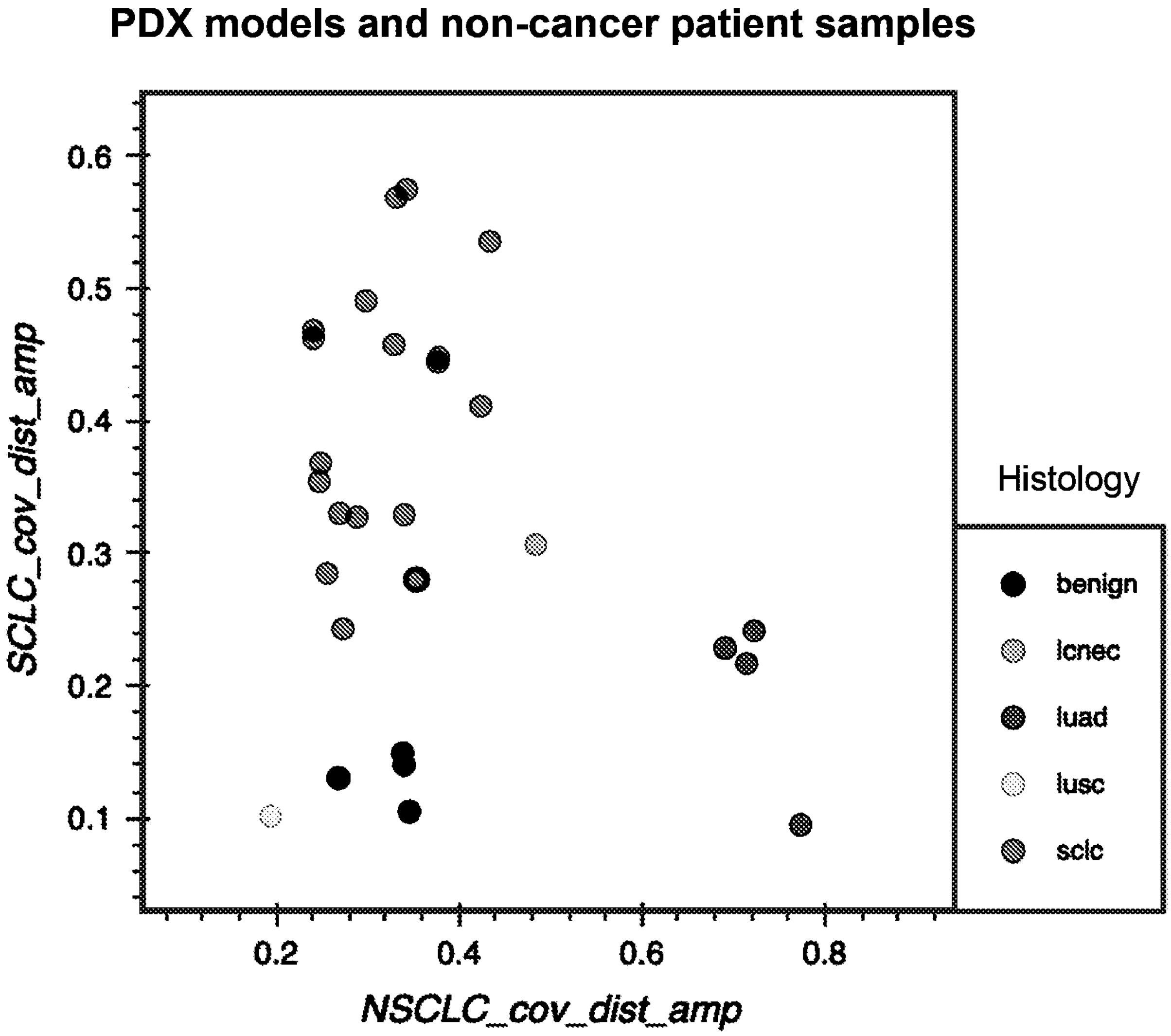

PDX models and non-cancer patient samples
SCLC_cov_dist_amp
0.6
0.5
0.4
0.3
0.2
0.1
0.2
0.4
0.6
0.8
NSCLC_cov_dist_amp
Histology
benign
lcnec
luad
lusc
sclc

FIG. 22A

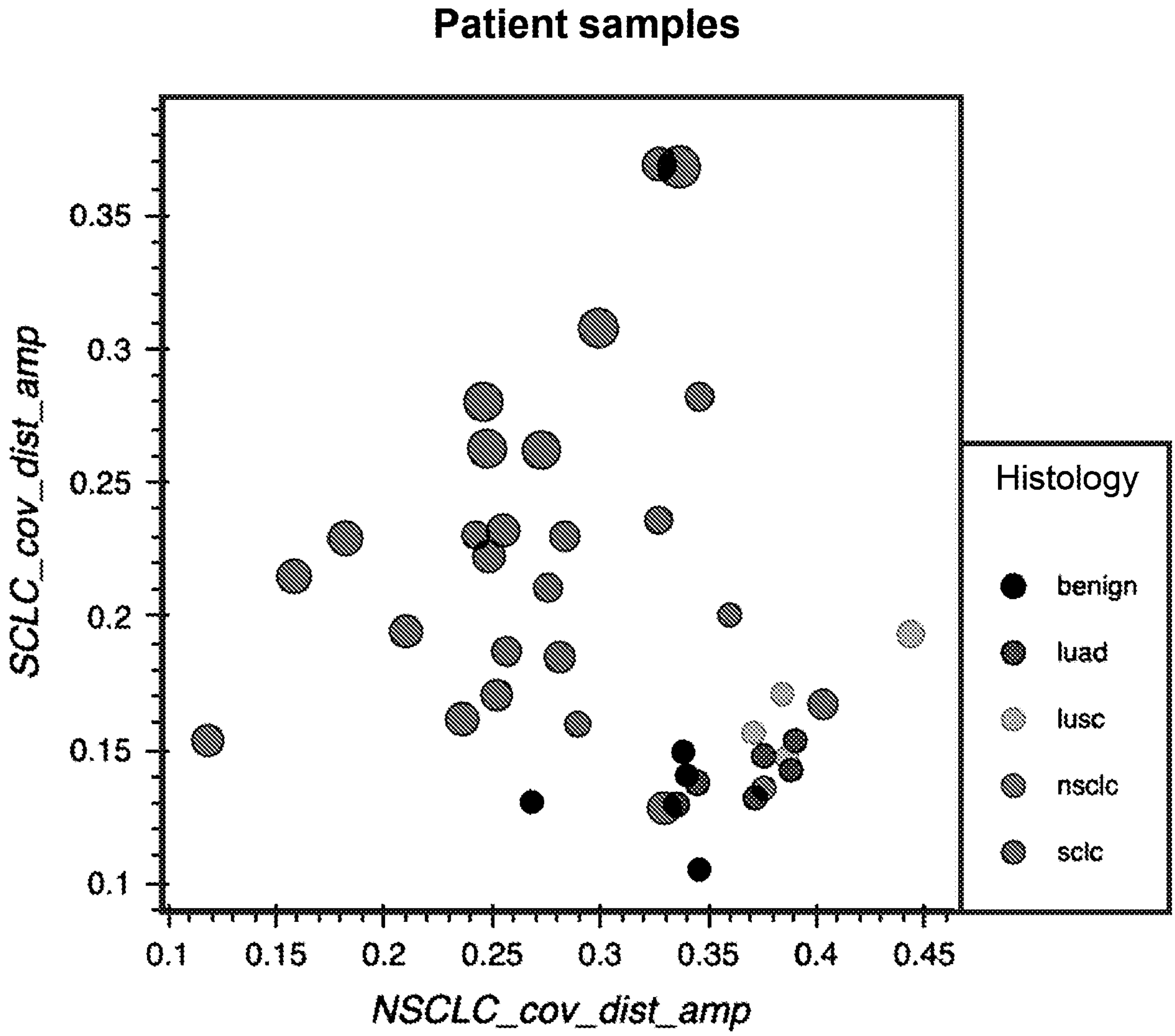

Patient samples
0.35
0.3
0.25
0.2
0.15
0.1
SCLC_cov_dist_amp
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
NSCLC_cov_dist_amp
Histology
benign
luad
lusc
nsclc
sclc

FIG. 22B

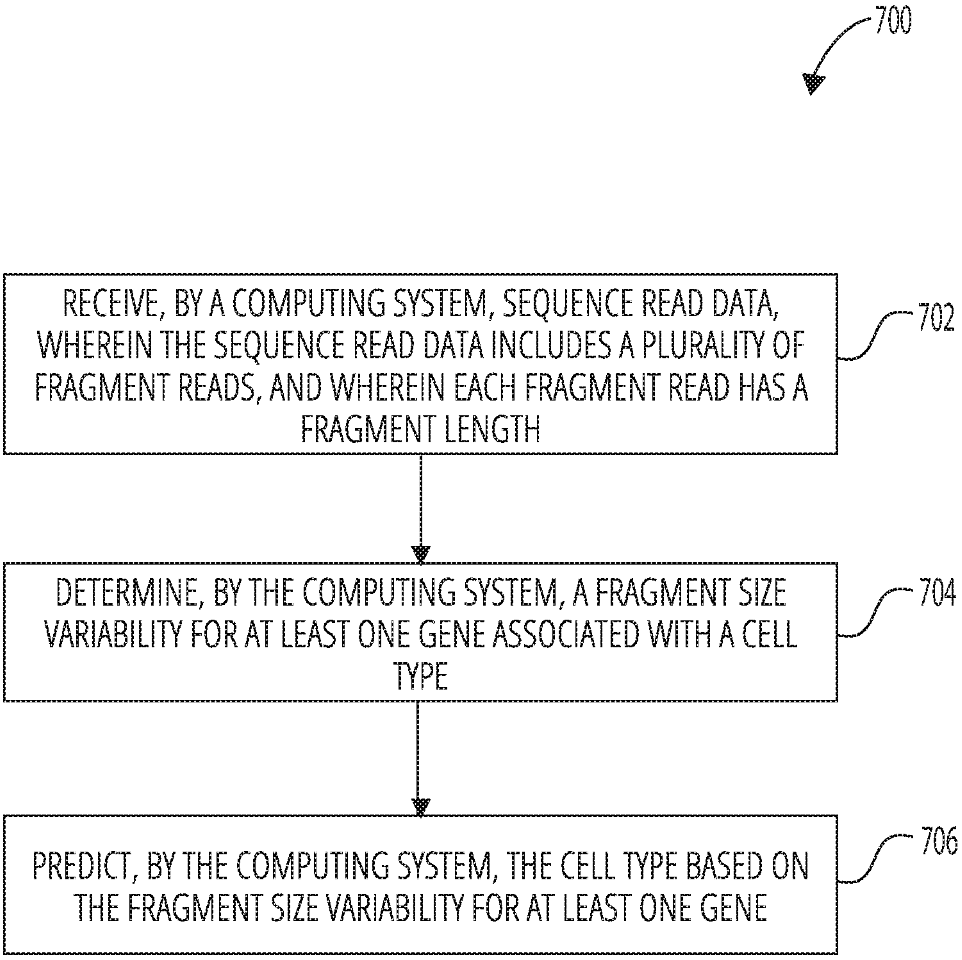

700
RECEIVE, BY A COMPUTING SYSTEM, SEQUENCE READ DATA, WHEREIN THE SEQUENCE READ DATA INCLUDES A PLURALITY OF FRAGMENT READS, AND WHEREIN EACH FRAGMENT READ HAS A FRAGMENT LENGTH
702
DETERMINE, BY THE COMPUTING SYSTEM, A FRAGMENT SIZE VARIABILITY FOR AT LEAST ONE GENE ASSOCIATED WITH A CELL TYPE
704
PREDICT, BY THE COMPUTING SYSTEM, THE CELL TYPE BASED ON THE FRAGMENT SIZE VARIABILITY FOR AT LEAST ONE GENE
706

*FIG. 23*