



(51) International Patent Classification:

G16B 20/10 (2019.01) G16B 20/20 (2019.01)

C12Q 1/6886 (2018.01) G16B 40/20 (2019.01)

G06F 17/18 (2006.01)

(21) International Application Number:

PCT/US2022/073167

(22) International Filing Date:

24 June 2022 (24.06.2022)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/215,281 25 June 2021 (25.06.2021) US

(71) Applicant: FOUNDATION MEDICINE, INC. [US/US];

150 Second Street, Cambridge, Massachusetts 02141 (US).

(72) Inventors: SOKOL, Ethan; c/o Foundation Medicine,

Inc., 150 Second Street, Cambridge, Massachusetts 02141

(US). MOORE, Jay; c/o Foundation Medicine, Inc., 150

Second Street, Cambridge, Massachusetts 02141 (US).

NEWBERG, Justin; c/o Foundation Medicine, Inc., 150

Second Street, Cambridge, Massachusetts 02141 (US). JIN,

Dexter; c/o Foundation Medicine, Inc., 150 Second Street,

Cambridge, Massachusetts 02141 (US). CHEN, Kuei-

Ting; c/o Foundation Medicine, Inc., 150 Second Street,

Cambridge, Massachusetts 02141 (US).

(74) Agent: WEINER, Ethan M. et al.; Morrison & Foerster

LLP, 755 Page Mill Road, Palo Alto, California 94304-1018

(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, IQ, 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,

(54) Title: SYSTEM AND METHOD OF CLASSIFYING HOMOLOGOUS REPAIR DEFICIENCY

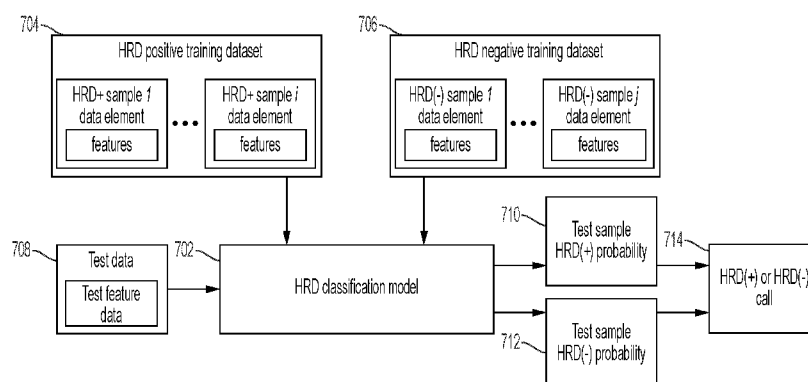


FIG. 7

(57) Abstract: Described herein are methods, devices, and systems for identifying a subset of a plurality of features, using one or more feature importance metrics, for training and using a homologous repair deficiency (HRD) classification model. Further described are methods, devices, and systems for classifying a tumor of a cancer, such as pancreatic cancer, as likely HRD positive or likely HRD negative, and for calling the tumor as HRD positive or HRD negative. Also described herein are methods of treating a tumor of a cancer, such as pancreatic cancer, based on the classifications

TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

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

**Declarations under Rule 4.17:**

— *of inventorship (Rule 4.17(iv))*

**Published:**

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

## **SYSTEM AND METHOD OF CLASSIFYING HOMOLOGOUS REPAIR DEFICIENCY**

### **CROSS-REFERENCE TO RELATED APPLICATIONS**

**[0001]** This application claims priority benefit to U.S. Provisional Application No. 63/215,281, filed on June 25, 2021, titled “SYSTEM AND METHOD OF CLASSIFYING HOMOLOGOUS REPAIR DEFICIENCY”, the contents of which are incorporated herein by reference for all purposes.

### **FIELD OF THE INVENTION**

**[0002]** Described herein are methods, devices, and systems for selecting features for a homologous repair deficiency (HRD) model, assessing tumors using the HRD model, and treating a tumor based on the assessment.

### **BACKGROUND OF THE INVENTION**

**[0003]** Copy number aberrations involve the deletion or amplification of large contiguous segments of the genome, and are common mutations in cancer. Certain copy number aberrations are associated with an inability to repair the genome by homologous recombination repair mechanisms, termed homologous repair deficiency (HRD). To identify some tumors with HRD, it is possible to sequence mutations in genes involved in the homologous repair pathway. Alternatively, it is possible to detect genomic scarring, which is the physical consequence of HRD, regardless of its cause.

**[0004]** Tumor genomes exhibiting HRD are associated with sensitivity to certain drugs, such as platinum chemotherapies or poly(ADP)-ribose polymerase (PARP) inhibitors. However, certain tumors remain difficult to classify as HRD positive. Thus, there remains a need to classify tumors of cancer, such as pancreatic, breast, or prostate cancer, where it is especially important, as HRD positive or HRD negative, so that appropriate treatments can be selected and administered to subjects. In the past, techniques for identifying HRD have suffered from inaccuracy and inefficiencies that have not allowed them to be used in practice. One reason for this is that feature selection techniques are currently insufficient to be able to accurately determine the HRD status of a sample in order to identify (e.g., classify) said tumors as HRD positive or HRD negative efficiently and accurately, e.g., due to overfitting. Another reason for this is that determining which features to identify to accurately determine the HRD status may also be a challenge. Accordingly, there is a need techniques and systems that accurately

and efficiently select a subset of features from a plurality of features that can be used train a model for performing said identification.

## SUMMARY OF THE INVENTION

**[0005]** Described herein are methods comprising: providing a genome obtained from a tumor of a subject; optionally, ligating one or more adapters onto the genome; amplifying nucleic acid molecules from the genome; capturing nucleic acid molecules from the amplified genome, wherein the captured nucleic acid molecules are captured by hybridization to one or more bait molecules; deriving, from the captured nucleic acid molecules, a set of input features; inputting, by one or more processors, the set of input features into a trained homologous recombination deficiency (HRD) model to identify the tumor as HRD-positive or HRD-negative using the trained HRD model, wherein the model is trained by: determining one or more feature importance metrics associated with each feature of a plurality of features, identifying a subset of features in the plurality of features using the one or more feature importance metrics, and training, by the one or more processors, the HRD model based on the identified subset of features; and classifying, by the one or more processors, using the trained HRD model, the tumor as HRD-positive or HRD-negative.

**[0006]** Further described herein are methods comprising: receiving, by one or more processors, a plurality of features; identifying, by the one or more processors, a subset of features in the plurality of features using one or more feature importance metrics; and training, by the one or more processors, a homologous recombination deficiency (HRD) model based on the identified subset of the plurality of features, wherein the HRD model is configured to receive sample data associated with a genome of a tumor in a subject and identify the tumor in the subject as HRD-positive or HRD-negative using the sample data.

**[0007]** Further described herein are methods comprising: receiving, by one or more processors, sample data associated with a genome of a tumor in a subject; inputting, by the one or more processors, the sample data into a trained homologous recombination deficiency (HRD) model, wherein the HRD model is trained by: determining one or more feature importance metrics associated with each feature of a plurality of features, identifying a subset of features in the plurality of features using the one or more feature importance metrics, and training, by the one or more processors, the HRD model based on the identified subset of features; and classifying, by the one or more processors, using the trained HRD model, the tumor as HRD-positive or HRD-negative.

**[0008]** In some embodiments of the described methods, the plurality of features comprises one or more copy number features, one or more short variant features, or a combination thereof. In some embodiments of the described methods, the one or more feature importance metrics comprise one or more of a Chi-Square test, analysis of variance (ANOVA), random forest, or gradient boosting.

**[0009]** In some embodiments of the described methods, identifying the subset of features in the plurality of features comprises: obtaining, by the one or more processors, one or more feature rankings according to the one or more feature importance metrics; and selecting, by the one or more processors, the subset of the plurality of features based on one or more feature rankings.

**[0010]** In some embodiments of the described methods, identifying the subset of the plurality of features comprises: (a) obtaining, by one or more processors, a feature ranking of the plurality of features according to a feature importance metric; (b) obtaining, by the one or more processors, a new feature set by adding one or more features from the plurality of features to an existing feature set based on the feature ranking; (c) training, by the one or more processors, a new HRD model using the new feature set; (d) evaluating, by the one or more processors, the trained new HRD model to obtain an evaluation result; and (e) storing, by the one or more processors, the evaluation result associated with the new HRD model and the new feature set; (f) repeating, by the one or more processors, steps (b)-(e) to obtain a plurality of evaluation results until a condition is met; and (g) selecting, by the one or more processors, the subset of the plurality of features based on the plurality of evaluation results.

**[0011]** In some embodiments of the described methods, the trained HRD model is a classification model, the method further comprising: receiving new sample data associated with a genome of a tumor in a new subject, wherein the new sample data is related to the subset of the plurality of features; providing the new sample data to the trained HRD classification model to produce a classification result of HRD-positive or HRD-negative; and outputting the classification result. In some embodiments, the classification result comprises at least one of a HRD-positive likelihood score and a HRD-negative likelihood score. In some embodiments, the method comprises recording, in a digital electronic file associated with the new subject, at least one of the HRD-positive likelihood score and the HRD-negative likelihood score. In some embodiments, the method comprises recording in a digital electronic file associated with the new subject that the tumor is HRD positive based on the HRD positive likelihood score or a designation that the tumor is HRD negative based on the HRD negative likelihood score.

**[0012]** In some embodiments of the described methods, the HRD model is a classification model, a regression model, a neural network, or any combination thereof. In some embodiments, the method comprises recording, in a digital electronic file associated with the new subject, at least one of the HRD-positive likelihood score and the HRD-negative likelihood score. In some embodiments, the method comprises recording in a digital electronic file associated with the new subject that the tumor is HRD positive based on the HRD positive likelihood score or a designation that the tumor is HRD negative based on the HRD negative likelihood score.

**[0013]** In some embodiments of the described methods, the plurality of features comprise at least one of a segment minor allele frequency (segMAF) feature, a number of sequencing reads feature, a segment size feature, a breakpoint count per  $x$  megabases feature, a change point copy number feature, a segment copy number feature, a breakpoint count per chromosome arm feature, or a number of segments with oscillating copy number feature. In some embodiments of the described methods, at least one of the plurality of features is assessed across the centromeric portion of the genome. In some embodiments of the described methods, at least one of the plurality of features is assessed across the telomeric portion of the genome.

**[0014]** In some embodiments of the described methods, at least one of the plurality of features is assessed across both the centromeric and telomeric portions of the genome.

**[0015]** In some embodiments of the described methods, the plurality of features comprise a breakpoint count per  $x$  megabases feature, wherein the breakpoint count per  $x$  megabases feature is based on the number of breakpoints appearing in windows of  $x$  megabases in length across the genome. In some embodiments, breakpoint count per  $x$  megabases feature is assessed across: (i) the telomeric portion of the genome; (ii) the centromeric portion of the genome; or (iii) both the telomeric portion and the centromeric portion of the genome. In some embodiments,  $x$  is between about 1 and about 100 megabases. In some embodiments,  $x$  is about 10 megabases, about 25 megabases, about 50 megabases, or about 100 megabases. In some embodiments, the breakpoint count per  $x$  megabases feature is a binned feature.

**[0016]** In some embodiments of the described methods, the plurality of features comprise a change point copy number feature, wherein the change point copy number is based on the absolute difference in copy number between adjacent genome segments across the genome of the tumor of the subject. In some embodiments, the change point copy number feature is derived from ploidy-normalized copy number data. In some embodiments, change point copy number feature is assessed across: (i) the telomeric portion of the genome; (ii) the

centromeric portion of the genome; or (iii) both the telomeric portion and the centromeric portion of the genome. In some embodiments, the change point copy number feature is a binned feature.

**[0017]** In some embodiments of the described methods, the plurality of features comprise a segment copy number feature, wherein segment copy number is based on the copy number of each genome segment. In some embodiments, the segment copy number feature is assessed across: (i) the telomeric portion of the genome; (ii) the centromeric portion of the genome; or (iii) both the telomeric portion and the centromeric portion of the genome. In some embodiments, the segment copy number feature is derived from ploidy-normalized copy number data. In some embodiments, the segment copy number feature is a binned feature.

**[0018]** In some embodiments of the described methods, the plurality of features comprise a breakpoint count per chromosome arm feature in the genome of the tumor of the subject. In some embodiments, the breakpoint count per chromosome arm feature is assessed across: (i) the telomeric portion of the genome; (ii) the centromeric portion of the genome; or (iii) both the telomeric portion and the centromeric portion of the genome. In some embodiments, the breakpoint count per chromosome arm feature is a binned feature.

**[0019]** In some embodiments of the described methods, the plurality of features comprise a number of segments with oscillating copy number feature. In some embodiments, the number of segments with oscillating copy number feature is based on the number of repeated alternating segments between two copy numbers across the genome of the tumor of the subject. In some embodiments, number of segments with oscillating copy number feature is assessed across: (i) the telomeric portion of the genome; (ii) the centromeric portion of the genome; or (iii) both the telomeric portion and the centromeric portion of the genome. In some embodiments, the number of segments with oscillating copy number feature is a binned feature.

**[0020]** In some embodiments of the described methods, the one or more copy number features comprise a segment minor allele frequency (segMAF) feature, wherein segMAF is based on the minor allele frequency at heterozygous single nucleotide polymorphisms. In some embodiments, segMAF is assessed across: (i) the telomeric portion of the genome; (ii) the centromeric portion of the genome; or (iii) both the telomeric portion and the centromeric portion of the genome. In some embodiments, the segMAF feature is a binned feature.

**[0021]** In some embodiments of the described methods, the one or more copy number features comprise a number of sequencing reads feature. In some embodiments, the number of sequencing reads feature is a binned feature.

[0022] In some embodiments of the described methods, the plurality of features further comprise a measure of genome-wide loss of heterozygosity of the genome of the tumor of the subject.

[0023] In some embodiments of the described methods, the plurality of features comprise one or more short variant features. In some embodiments, the one or more short variant features comprise at least one of a deletions in microhomology or repetitive regions feature and a mutational signature derived from two or more short variant features. In some embodiments, the deletions in microhomology or repetitive regions feature are deletions of at least 5 basepairs.

[0024] In some embodiments of the described methods, training the HRD model comprises: receiving, by the one or more processors, an HRD-positive training dataset, wherein the HRD-positive training dataset comprises a plurality of features associated with an HRD-positive tumor and an HRD-positive label; receiving, by the one or more processors, an HRD-negative training dataset, wherein the HRD-negative training dataset comprises a plurality of features associated with an HRD-negative tumor and an HRD-negative label; training, by the one or more processors, the HRD model using the HRD-positive training dataset and the HRD-negative training dataset. In some embodiments, training comprises using a HRD-positive training dataset and an HRD-negative training dataset. In some embodiments, the method comprises balancing, by the one or more processors, the HRD-positive training dataset and the HRD-negative training dataset prior to training the HRD model.

[0025] In some embodiments of the described methods, the method further comprises testing, by the one or more processors, the trained model using a HRD-positive testing dataset comprising a HRD-positive control derived from a genome sequence comprising loss-of-function mutations in *BRCA1*, *BRCA2*, both *BRCA1* and *BRCA2*, or biallelic mutations of *BRCA1* and *BRCA2*. In some embodiments, training comprises using a HRD-positive training dataset and an HRD-negative training dataset. In some embodiments, the method comprises balancing, by the one or more processors, the HRD-positive training dataset and the HRD-negative training dataset prior to training the HRD model.

[0026] In some embodiments of the described methods, the method further comprises testing, by the one or more processors, the trained model using a HRD-positive testing dataset comprising a HRD-positive control derived from a genome sequence comprising loss-of-function mutations in at least one of *ATM*, *BARD1*, *BRIP1*, *CDK12*, *CHEK1*, *CHEK2*, *FANCL*, *PALB2*, *RAD51B*, *RAD51C*, *RAD51D*, or *RAD45L*. In some embodiments, training

comprises using a HRD-positive training dataset and an HRD-negative training dataset. In some embodiments, the method comprises balancing, by the one or more processors, the HRD-positive training dataset and the HRD-negative training dataset prior to training the HRD model.

**[0027]** In some embodiments of the described methods, the method further comprises testing, by the one or more processors, the trained model using a HRD-negative testing dataset comprising a HRD-negative training dataset comprising a HRD-negative control derived from a consensus human genome sequence. In some embodiments, training comprises using a HRD-positive training dataset and an HRD-negative training dataset. In some embodiments, the method comprises balancing, by the one or more processors, the HRD-positive training dataset and the HRD-negative training dataset prior to training the HRD model.

**[0028]** In some embodiments of the described methods, the tumor in the subject is a prostate cancer, non-small cell lung cancer (NSCLC), colorectal cancer (CRC), ovarian cancer, breast cancer, or pancreatic cancer.

**[0029]** In some embodiments of the described methods, training the HRD model comprises fitting the HRD model to sample data associated with ovarian cancer, non-small cell lung cancer (NSCLC), colorectal cancer (CRC), breast cancer, pancreatic cancer, or prostate cancer, wherein the sample data comprises the subset of the plurality of features.

**[0030]** In some embodiments of the described methods, the tumor is obtained from a sample that is a solid tissue biopsy sample. In some embodiments, the solid tissue biopsy sample is a formalin-fixed paraffin-embedded (FFPE) sample. In some embodiments of the described methods, the tumor is obtained from a sample that is a liquid biopsy sample comprising circulating tumor DNA (ctDNA). In some embodiments of the described methods, the tumor is obtained from a sample that is a liquid biopsy sample comprising cell-free DNA (cfDNA).

**[0031]** In some embodiments of the described methods, the method further comprises: determining, identifying, or applying the output of the tumor as HRD-positive or HRD-negative as a diagnostic value associated with the patient. In some embodiments of the described methods, the method further comprises generating a genomic profile for the subject based on the output of the tumor as HRD-positive or HRD-negative. In some embodiments, the method further comprises administering an anti-cancer agent or applying an anti-cancer treatment to the subject based on the generated genomic profile. In some embodiments of the described methods, the output of the tumor as HRD-positive or HRD-negative is used in generating a genomic profile for the subject. In some embodiments of the described methods, the output of the tumor as HRD-positive or HRD-negative is used in making suggested

treatment decisions for the subject. In some embodiments of the described methods, the output of the tumor as HRD-positive or HRD-negative is used in applying or administering a treatment to the subject.

**[0032]** In some embodiments of the described methods, the HRD model is a machine learning model.

**[0033]** In some embodiments of the described methods, the subject has a cancer, is at risk of having a cancer, or is suspected of having a cancer.

**[0034]** Further described herein are methods of treating cancer in a subject, comprising: (a) identifying the tumor as HRD-positive or HRD-negative according to any method described above; (b) administering to the subject a therapeutically effective amount of a drug effective in a HRD positive tumor if the tumor of the cancer is assessed as HRD positive. In some embodiments, the drug effective in a HRD positive tumor is a platinum-based drug or a PARP inhibitor. In some embodiments, the method comprises administering to the subject a therapeutically effective amount of a drug that is not a platinum-based drug or a PARP inhibitor if the tumor is assessed as HRD negative.

**[0035]** Further described herein are methods for selecting a therapy for a cancer in a subject, the method comprising: (a) assessing a tumor of the cancer as HRD-positive or HRD-negative according to any method described above; (b) selecting a therapy that is effective in a HRD positive tumor if the cancer is assessed as HRD positive. In some embodiments, the method comprises selecting a therapy that is not a platinum-based drug or a PARP inhibitor if the tumor is assessed as HRD negative. In some embodiments, the therapy that is effective in a HRD positive tumor is a platinum-based drug or a PARP inhibitor.

**[0036]** Further described herein are computer systems, comprising: one or more processors; a memory; and one or more programs, wherein the one or more programs are stored in the memory and configured to be executed by the one or more processors, the one or more programs including instructions for: performing any one of the methods described above.

**[0037]** Further described herein are non-transitory computer-readable storage media storing one or more programs, the one or more programs comprising instructions, which when executed by one or more processors of an electronic device, cause the electronic device to perform the any one of the methods described above.

## BRIEF DESCRIPTION OF THE DRAWINGS

[0038] **FIG. 1** shows an exemplary process for classifying a tumor of a cancer in a subject as HRD positive (HRD(+)) or HRD negative (HRD(-)).

[0039] **FIG. 2** shows different types of features that may be evaluated using different feature importance metrics such as ANOVA, random forest, gradient boosting (e.g., XGB), and Chi-Squared.

[0040] **FIG. 3A** shows an exemplary feature overlap analysis.

[0041] **FIG. 3B** shows an exemplary feature overlap analysis.

[0042] **FIG. 4** shows an exemplary iterative feature selection process.

[0043] **FIG. 5** shows an example plot of model performances obtained from an exemplary iterative feature selection process.

[0044] **FIG. 6A** shows an exemplary cross-validation process which may be used to evaluate and tune the performance of a model.

[0045] **FIG. 6B** shows an exemplary division of a plurality of data elements into equally-sized subsets.

[0046] **FIG. 7** shows an exemplary method for training and operating the HRD classification model configured to classify a tumor of a cancer in a subject as HRD positive (HRD(+)) or HRD negative (HRD(-)).

[0047] **FIG. 8** shows an example of HRD score distributions for different machine learning models using logistic regression, gradient boosting (e.g., XGB), and random forest.

[0048] **FIG. 9** shows an example model performance in samples stratified by HRD and/or *BRCA1/2* mutation status. The left side shows the pool of sample tumors designated “HRD WildType: True” (N=245,050; -1 on the right side of figure), “HRD WildType:False” (N=30,799; 0 on right side of figure) and true HRD-positive samples (biallelic *BRCA* mutation; N=6,851; 1 on right side of figure).

[0049] **FIG. 10** shows the example model performance from the subsets of **FIG. 9** in different tumor types (breast, ovarian, pancreatic, and prostate cancer). For each tumor type, the subsets correspond to the subsets -1, 0, and 1 of **FIG. 9** (i.e., HRD WildType: True, HRD WildType: False, and biallelic *BRCA* mutation for each cancer, respectively).

[0050] **FIG. 11** shows an example of a computing device in accordance with one embodiment, which may be used with certain methods described herein.

## DETAILED DESCRIPTION OF THE INVENTION

**[0051]** Described herein are computer-implemented methods of identifying a subset of a plurality of features using one or more feature importance metrics for training a homologous recombination deficiency (HRD) model (e.g., a classification model). The model is configured to receive test sample data related to the subset of the plurality of features associated with a genome of a tumor in a subject and identify (e.g., classify) the tumor as likely HRD positive or likely HRD negative. Further described herein are methods of identifying (e.g., classifying) a tumor, such as a prostate cancer, ovarian cancer, breast cancer, colorectal cancer, NSCLC, or pancreatic cancer tumor, as likely HRD positive (HRD(+)) or likely HRD negative (HRD(-)). Further described herein are methods of treating a cancer, such as, but not limited to, pancreatic, prostate, ovarian, breast cancer, non-small cell lung cancer (NSCLC), or colorectal cancer (CRC), based on the identification of a tumor as HRD positive (or likely HRD positive) or HRD negative (or likely HRD negative).

**[0052]** Selecting a subset of features can reduce overfitting of the model. Overfitting is problematic because it reduces the scalability of the model and can result in inaccurate classifications (e.g., inaccurate HRD status) because the model ignores scenarios that fall outside of the data used to train the model. Further, by selecting a subset of features that have higher feature importance, the classification model can be trained with less training data and would require less input data. This not only allows for a more efficient modeling process, but also a more accurate classification from a broader range of samples from the model. Further, a model with a reduced set of input features can require less processing power for training and for performing the classification task. Thus, the feature selection process improves the functioning of a computer system by improving processing speed and allowing for efficient use of computer memory and processing power. In addition, by selecting from certain derived copy number features and/or short variant features, the trained model provides greater efficiency and accuracy (e.g., less false-positives/false-negatives) when identifying tumors as HRD-positive or HRD-negative in comparison with previous methods. Previous methods of assessing HRD, such as loss of heterozygosity, telomeric allelic imbalance, and large-scale transition, are subject to noise and error compared with the assessment of derived copy number features and/or short variant features described herein. Proper identification of tumors is integral to being able to appropriately select a treatment for the patient (subject).

**[0053]** Oncogenesis is driven, in part, by the accumulation of somatic alterations of the genomes of cells. Among these alterations include copy number alterations, which are

common in many cancers. Loss-of-function, gain-of-function, or gene regulation mutations in certain genes involved in the homologous repair deficiency pathway can lead to accumulation of these copy number alterations. However, other than mutations in certain key genes, such as *BRCA1* and *BRCA2*, the precise combinations of mutations leading to HRD-positive status are unknown. Some tumors will be HRD positive through non-genomic means, for example, through promoter methylation of HRD-associated genes such as *BRCA1*. Instead of sequencing HRD-associated genes, an alternative approach is to identify and assess the consequences of HRD, such as changes in certain copy number features or in loss of heterozygosity features. However, while both HRD positive and HRD negative genomes may exhibit copy number alterations, the precise values and combinations of features that indicate the presence of HRD are unknown.

**[0054]** Thus, in one aspect, the methods of the invention relate to selecting a subset of features (from a larger plurality of potential features) that can be used to train and operate an HRD classifier process. In another aspect, the methods of the invention relate generally to means of identifying (e.g., classifying) tumors as likely HRD positive (HRD(+)) or likely HRD negative (HRD(-)) based, at least in part, on assessments of features, such as features corresponding to copy number aberrations. This classification is generally based on an assessment of the likelihood that the tumor is HRD-positive or HRD-negative. Based on this assessment, the HRD classifier process may further call the tumor as HRD positive or HRD negative. This classification and/or call may be used as a diagnostic value for the patient having the tumor.

**[0055]** Existing methods for classifying tumors as likely HRD positive or likely HRD negative are often unreliable or imprecise, particularly for HRD positive tumors having wild-type *BRCA1* and *BRCA2* (which are sometimes described as tumors having a “*BRCAness*” profile, i.e., those tumors which exhibit similarities to *BRCA1/2*-mutant tumors without having the associated *BRCA1/2* mutations). Alternatively, not all mutations, even pathogenic mutations such as *BRCA1/2* alterations, result in HRD (e.g., some mutations may be monoallelic passengers). Homologous repair deficiency associated with cancer scars the tumor cell genome leading to detectable changes in copy number (i.e., copy number aberrations) and/or indel patterns. The particular pattern, distribution, and form of these copy number aberrations and/or indel patterns can be used to classify tumors into HRD phenotype classes. The present application, in various embodiments, provides means to select the features associated with these patterns (i.e., copy number features) and indel patterns (i.e.,

short variant features) among other potential features (such as basic features as otherwise described herein) which can be used to identify HRD-positive tumors.

**[0056]** The present application further provides specifically configured models that are based on one or more data features (such as one or more copy number features and/or one or more short variant features) associated with a genome of a cancerous tumor in a subject which can more reliably identify (e.g., classify) said tumors as likely HRD positive or likely HRD negative and optionally call the tumors as HRD positive or HRD negative. The identification (e.g., classification) of a tumor of a cancer in a subject indicates how the tumor should be treated. A trained HRD model using test data comprising at least one or more copy number features, including, for example, one or more of a segment size feature, a sequencing reads feature, an absolute copy number feature, a breakpoint count per  $x$  megabases feature, a change point copy number feature, a segment copy number feature, a breakpoint count per chromosome arm feature, a number of segments with oscillating copy number feature, and a segment minor allele frequency feature can be used to identify (e.g., classify) a test tumor as likely HRD positive or likely HRD negative, and also call the tumor as HRD positive or HRD negative based on the likelihood score. These categories of copy number features have been identified as being useful for this identification. Certain categories of short variant features have also been identified as being useful for this identification, including, but not limited to, a deletions (e.g., of at least 5-basepairs) in, for example, microhomology or repetitive regions feature and/or a mutational signature incorporating two or more short variant features.

**[0057]** In combination with one or more of these copy number features and/or one or more of these short variant features, other features or measures may be useful in the described methods, including, but not limited to, certain basic features such as age of subject, cancer type, cancer stage, tumor purity, tumor genome ploidy, and/or tumor genome loss of heterozygosity.

**[0058]** Once a tumor of a cancer in a subject has been identified (e.g., classified) as likely HRD positive or likely HRD negative, or called as HRD positive or HRD negative, it may be treated with an appropriate therapy. For example, if the tumor is identified as likely HRD positive, it may be treated with a drug effective in a HRD positive cancer, such as a platinum-based drug or a PARP inhibitor.

## **Definitions**

**[0059]** As used herein, the singular forms “a,” “an,” and “the” include the plural reference unless the context clearly dictates otherwise.

**[0060]** Reference to “about” a value or parameter herein includes (and describes) variations that are directed to that value or parameter per se. For example, description referring to “about X” includes description of “X”.

**[0061]** The terms "cancer" and "cancerous" refer to or describe the physiological condition in mammals that is typically characterized by unregulated cell growth. Included in this definition are benign and malignant cancers. By "early stage cancer" or "early stage tumor" is meant a cancer that is not invasive or metastatic or is classified as a Stage 0, 1, or 2 cancer. Examples of a cancer include, but are not limited to, a lung cancer (e.g., a non-small cell lung cancer (NSCLC)), a kidney cancer (e.g., a kidney urothelial carcinoma), a bladder cancer (e.g., a bladder urothelial (transitional cell) carcinoma), a breast cancer, a colorectal cancer (e.g., a colon adenocarcinoma), an ovarian cancer, a pancreatic cancer, a gastric carcinoma, an esophageal cancer, a mesothelioma, a melanoma (e.g., a skin melanoma), a head and neck cancer (e.g., a head and neck squamous cell carcinoma (HNSCC)), a thyroid cancer, a sarcoma (e.g., a soft-tissue sarcoma, a fibrosarcoma, a myxosarcoma, a liposarcoma, an osteogenic sarcoma, an osteosarcoma, a chondrosarcoma, an angiosarcoma, an endotheliosarcoma, a lymphangiosarcoma, a lymphangioendotheliosarcoma, a leiomyosarcoma, or a rhabdomyosarcoma), a prostate cancer, a glioblastoma, a cervical cancer, a thymic carcinoma, a leukemia (e.g., an acute lymphocytic leukemia (ALL), an acute myelocytic leukemia (AML), a chronic myelocytic leukemia (CML), a chronic eosinophilic leukemia, or a chronic lymphocytic leukemia (CLL)), a lymphoma (e.g., a Hodgkin lymphoma or a non-Hodgkin lymphoma (NHL)), a myeloma (e.g., a multiple myeloma (MM)), a mycoses fungoides, a merkel cell cancer, a hematologic malignancy, a cancer of hematological tissues, a B cell cancer, a bronchus cancer, a stomach cancer, a brain or central nervous system cancer, a peripheral nervous system cancer, a uterine or endometrial cancer, a cancer of the oral cavity or pharynx, a liver cancer, a testicular cancer, a biliary tract cancer, a small bowel or appendix cancer, a salivary gland cancer, an adrenal gland cancer, an adenocarcinoma, an inflammatory myofibroblastic tumor, a gastrointestinal stromal tumor (GIST), a colon cancer, a myelodysplastic syndrome (MDS), a myeloproliferative disorder (MPD), a polycythemia Vera, a chordoma, a synovioma, an Ewing's tumor, a squamous cell carcinoma, a basal cell carcinoma, an adenocarcinoma, a sweat gland carcinoma, a sebaceous gland carcinoma, a papillary carcinoma, a papillary adenocarcinoma, a medullary carcinoma, a bronchogenic carcinoma, a renal cell carcinoma, a hepatoma, a bile duct carcinoma, a choriocarcinoma, a seminoma, an embryonal carcinoma, a Wilms' tumor, a bladder carcinoma, an epithelial carcinoma, a glioma, an astrocytoma, a medulloblastoma, a

craniopharyngioma, an ependymoma, a pinealoma, a hemangioblastoma, an acoustic neuroma, an oligodendroglioma, a meningioma, a neuroblastoma, a retinoblastoma, a follicular lymphoma, a diffuse large B-cell lymphoma, a mantle cell lymphoma, a hepatocellular carcinoma, a thyroid cancer, a small cell cancer, an essential thrombocythemia, an agnogenic myeloid metaplasia, a hypereosinophilic syndrome, a systemic mastocytosis, a familial hypereosinophilia, a neuroendocrine cancer, or a carcinoid tumor.

**[0062]** The tumor “tumor,” as used herein, refers to all neoplastic cell growth and proliferation, whether malignant or benign, and all pre-cancerous and cancerous cells and tissues. The terms “cancer,” “cancerous,” and “tumor” are not mutually exclusive as referred to herein.

**[0063]** The terms “individual,” “patient,” and “subject” are used synonymously, and refer to a mammal, and includes, but is not limited to, human, bovine, horse, feline, canine, rodent, or primate. In one embodiment, the subject is a human.

**[0064]** The terms “effective amount” or “therapeutically effective amount” as used herein refer to an amount of a compound, drug, or composition sufficient to treat a specified disorder, condition or disease, such as ameliorate, palliate, lessen, and/or delay one or more of its symptoms. In reference to a cancer, an effective amount comprises an amount sufficient to cause the number of cancer cells present in a subject to decrease in number and/or size and/or to slow the growth rate of the cancer cells. In some embodiments, an effective amount is an amount sufficient to prevent or delay recurrence of the disease. In the case of cancer, the effective amount of the compound or composition may: (i) reduce the number of cancer cells; (ii) inhibit, retard, slow to some extent and preferably stop cancer cell proliferation; (iii) prevent or delay occurrence and/or recurrence of the cancer; and/or (iv) relieve to some extent one or more of the symptoms associated with the cancer.

**[0065]** As used herein, “treatment” or “treating” is an approach for obtaining beneficial or desired results including clinical results. For purposes of this invention, beneficial or desired clinical results include, but are not limited to, one or more of the following: alleviating one or more symptoms resulting from the disease, diminishing the extent of the disease, stabilizing the disease (*e.g.*, preventing or delaying the worsening of the disease), preventing or delaying the spread (*e.g.*, metastasis) of the disease, preventing or delaying the recurrence of the disease, delay or slowing the progression of the disease, ameliorating the disease state, providing a remission (partial or total) of the disease, decreasing the dose of one or more other medications required to treat the disease, delaying the progression of the disease,

increasing the quality of life, and/or prolonging survival. In reference to a cancer, the number of cancer cells present in a subject may decrease in number and/or size and/or the growth rate of the cancer cells may slow. In some embodiments, treatment may prevent or delay recurrence of the disease. In the case of cancer, the treatment may: (i) reduce the number of cancer cells; (ii) inhibit, retard, slow to some extent and preferably stop cancer cell proliferation; (iii) prevent or delay occurrence and/or recurrence of the cancer; and/or (iv) relieve to some extent one or more of the symptoms associated with the cancer. The methods of the invention contemplate any one or more of these aspects of treatment.

**[0066]** It is understood that aspects and variations of the invention described herein include “consisting” and/or “consisting essentially of” aspects and variations.

**[0067]** When a range of values is provided, it is to be understood that each intervening value between the upper and lower limit of that range, and any other stated or intervening value in that states range, is encompassed within the scope of the present disclosure. Where the stated range includes upper or lower limits, ranges excluding either of those included limits are also included in the present disclosure.

**[0068]** The section headings used herein are for organization purposes only and are not to be construed as limiting the subject matter described. The description is presented to enable one of ordinary skill in the art to make and use the invention and is provided in the context of a patent application and its requirements. Various modifications to the described embodiments will be readily apparent to those persons skilled in the art and the generic principles herein may be applied to other embodiments. Thus, the present invention is not intended to be limited to the embodiment shown but is to be accorded the widest scope consistent with the principles and features described herein.

**[0069]** The figures illustrate processes according to various embodiments. In the exemplary processes, some blocks are, optionally, combined, the order of some blocks is, optionally, changed, and some blocks are, optionally, omitted. In some examples, additional steps may be performed in combination with the exemplary processes. Accordingly, the operations as illustrated (and described in greater detail below) are exemplary by nature and, as such, should not be viewed as limiting.

**[0070]** The disclosures of all publications, patents, and patent applications referred to herein are each hereby incorporated by reference in their entireties. To the extent that any reference incorporated by reference conflicts with the instant disclosure, the instant disclosure shall control.

**Feature selection**

**[0071]** Starting with a plurality of features, including those as described otherwise herein, a subset of the plurality of features may be identified using one or more feature importance metrics. Generally, the feature importance metrics allow for evaluation of individual features to determine which features may be most relevant for assessing HRD. Exemplary feature importance metrics include, but are not limited to, gradient boosting (such as XGBoost, also known as XGB), analysis of variance (ANOVA), Chi-Squared analysis, and random forest. Individual features can be assigned values based on these feature importance metrics, where features are assigned increasing importance based on increasing contribution to the performance of the HRD model (e.g., improving performance of the model in classifying tumors as HRD-positive or HRD-negative). Features of higher importance, such as features above a threshold (such as features above median among the plurality of features) may then be selected for use in training or running the HRD model. Once the subset of features is identified, a HRD model (e.g., a classification model) may be trained using the subset of features. The HRD model may then be used to identify (e.g., classify) a tumor of a subject using test data obtained from the tumor and including at least a portion of the features identified during the feature selection.

**[0072]** By selecting this subset of features that have higher feature importance, the model can be trained with less training data and requires less input data, thus improving memory usage and management. Further, a model with a reduced set of input features requires less processing power for training and for performing the identification (e.g., classification) task. Thus, the feature selection process improves the functioning of a computer system by improving processing speed and allowing for efficient use of computer memory and processing power.

**[0073]** **FIG. 1** illustrates an exemplary process for classifying a tumor of a cancer in a subject as HRD-positive or HRD-negative including blocks for identifying a subset of a plurality of features, in accordance with some embodiments. In some embodiments, process 100 is performed, for example, using one or more electronic devices implementing a software platform. In some examples, process 100 is performed using a client-server system, and the blocks of process 100 are divided up in any manner between the server and client device(s). In other examples, process 100 is performed using only a client device or only multiple client devices. In process 100, some blocks are, optionally, combined, the order of some blocks is, optionally, changed, and some blocks are, optionally, omitted. In some examples, additional steps may be performed in combination with the process 100. Accordingly, the operations as

illustrated (and described in greater detail below) are exemplary by nature and, as such, should not be viewed as limiting.

**[0074]** At block 102 of **FIG. 1**, an exemplary system (e.g., one or more electronic devices) receives a plurality of features. In some embodiments, the system receives a dataset comprising a plurality of data elements. A data element can comprise data related to a plurality of features and an associated classification label (e.g. HRD-positive or HRD-negative). For example, a data element can comprise data related to the plurality of features of a sample from a particular subject, and an associated classification label indicating whether the sample is HRD-positive and HRD-negative. The features may include features categorized as basic features, copy number features, and/or short variant features (e.g., a feature corresponding to a base substitution or an indel (insertion or deletion)). Basic features may include, but are not limited to, features related to age of the patient from which the data were obtained, cancer type, cancer stage, tumor purity, tumor genome ploidy, and tumor genome loss of heterozygosity (such as percent of genome under loss of heterozygosity). Copy number features may include, but are not limited to, a segment size feature, a number of sequencing reads feature, an absolute copy number feature, a breakpoint count per  $x$  megabases feature, a change point copy number feature, a segment copy number feature, a breakpoint count per chromosome arm feature, a number of segments with oscillating copy number feature, and a segment minor allele frequency feature. Short variant features may include, but are not limited to, a deletions (for example, of at least 5-basepairs) in, for example, homopolymer or repetitive regions feature and/or a mutational signature incorporating two or more short variant features. In some embodiments, one or more of the features are binned features, wherein the values are sorted into bins, such as a binary, a tertile, a quartile, a quintile, a sextile, a septile, or any other suitable binning organization.

**[0075]** At block 104 of **FIG. 1**, the system and method selects a subset of features from the plurality of features (i.e., the basic features, the copy number features, and/or the short variant features). The subset of features selected may have relatively high predictive value for classifying a tumor of a cancer in a subject as HRD-positive or HRD-negative. In some embodiments, features that have relatively low predictive value and/or are redundant can be excluded from the subset of features in block 104. In some embodiments, the predictive value of a feature may be quantified using a feature importance metric. In some embodiments, the feature importance metric can be applied to obtain a feature importance score for each feature of the plurality of features. The feature importance score of a feature is obtained from a statistical correlation between the feature and the classification label (e.g., HRD-positive or

HRD-negative). The statistical correlation between the feature and the classification label may be interpreted based on how much predictive value the feature has for the classification task. In other words, a higher feature importance score can be achieved by having, for example, a higher statistical correlation between the feature and the classification label, which can indicate that the feature plays a more important role in predicting the classification label. By using features that have higher feature importance, a classification model can be trained with less data, thus providing a great degree of efficacy to the training process and less constraints on computer resources (e.g., memory usage, processing speed, etc.). For example, a model with a reduced set of input features can require fewer processing resources to train and perform the classification task. Finally, a model with a reduced set of input features may exhibit less noise and avoid overtraining. Thus, the feature selection process improves the functioning of a computer system by improving the overall efficacy of the training process, improving processing speed, and allowing for efficient use of computer memory and processing resources.

**[0076]** In some embodiments, the system selects the subset of features from the plurality of features received at block 102 of **FIG. 1** by performing a feature overlap analysis, as shown by block 104a. At block 104a, each feature importance metric is used to calculate feature importance scores of the plurality of features received from block 102. For each feature importance metric, the system can rank the plurality of features according to their feature importance scores. Thus, the system can obtain a plurality of feature rankings corresponding to the plurality of feature importance features. The system may then identify a subset of features based on the plurality of rankings. The process of ranking the features and identifying the subset of features is described in more detail below.

**[0077]** In some embodiments, different types of features can be evaluated using different feature importance metrics. **FIG. 2** illustrates a plurality of feature importance metrics that may be used to rank the plurality of features in block 104a in accordance with some embodiments. The depicted exemplary feature importance metrics include ANOVA, random forest, gradient boosting (e.g., XGB), and Chi-Squared. Further, ANOVA can be used to evaluate numeric features of the plurality of features to provide a ranking of the numeric features. Chi-Squared can be used to evaluate categorical features of the plurality of features to provide a ranking of the categorical features. Random forest can be used to evaluate all of the plurality of features to rank all features. Similarly, gradient boosting (such as XGB) can be used to evaluate all of the plurality of features to rank all features.

[0078] In some embodiments, the feature importance metrics comprise an analysis of variance (ANOVA) model. ANOVA assesses if there is equal variance between groups (i.e., HRD-positive or HRD-negative) when numeric input variables are compared to a classification target variable. If there is equal variance between groups, then the feature has no impact on the response and it may not be considered for model training. Based on the variance value (f-value), the features may be ranked, and those features that are, for example, above median may be selected as useful features for the model.

[0079] In some embodiments, the feature importance metrics comprise a Chi-Square analysis. For feature selection, Chi-Square analysis tests how expected count (i.e., if the feature is independent of output) and observed count deviate from each other. A higher Chi-Square value for a feature indicates it is more dependent on the response variables and is thus more important. Using Chi-Square analysis, features may be ranked, and those features that are, for example, above median may be selected as useful features for the model.

[0080] In some embodiments, the feature importance metrics comprise a random forest analysis. During feature selection, for each tree, the prediction accuracy on the out-of-bag portion of the data is recorded. The process is repeated after permuting each predictor variable. The difference between the two accuracies is then averaged over all trees, and normalized by the standard error.

[0081] In some embodiments, the feature importance metrics comprise a gradient boosting analysis (e.g., an extreme gradient boosting (XGB) analysis). Gradient boosting, such as XGB, tests the gain contribution of each feature to the model. For a boosted tree model, each gain of each feature of each tree is accounted for, and then the average per feature contribution is assessed. The highest percentage contributor features may then be selected.

[0082] At block 104a of **FIG. 1**, after the plurality of features are ranked according to feature importance metrics, the system uses the plurality of rankings to select a subset of features. An exemplary process of selecting a subset of features is described in further detail below in **FIGs. 3A** and **3B**.

[0083] **FIG. 3A** illustrates an exemplary feature overlap analysis in accordance with some embodiments. As described above in **FIG. 2**, a plurality of feature importance metrics may be used to rank a plurality of features. In the example of **FIG. 3A**, the exemplary process uses an ANOVA, a random forest, and a gradient boosting analysis to rank the features. However, those skilled in the art will understand that other learning techniques known in the art could be used as well. However, for exemplary purposes in **FIG. 3A**, the ANOVA feature ranking 302 includes features 1, 4, 5, and 8 as the highest ranking features; the random forest ranking

304 includes features 8, 2, 3, and 1 as the highest ranking features; the gradient boosting ranking 306 includes features 6, 1, 4, and 2 as the highest ranking features. In some embodiments, other feature importance metrics may be used to evaluate the features. In some embodiments, fewer or more than three metrics may be used to evaluate the features. In some embodiments, more than four features may be considered as high-ranking features, such as any of more than five, more than six, more than seven, more than eight, more than nine, more than ten, more than eleven, more than twelve, more than thirteen, more than fourteen, more than fifteen, more than sixteen, more than seventeen, more than eighteen, more than nineteen, more than twenty, more than twenty-one, more than twenty-two, more than twenty-three, more than twenty-four, or more than twenty-five features may be considered as high-ranking features.

**[0084]** Once the features have been ranked, the system may perform the feature overlap analysis to determine features that one or more metrics have identified as high-ranking features. In the example of **FIG. 3A**, feature overlap analysis 308 identifies feature 1 as a high-ranking feature identified in ANOVA feature ranking 302, random forest ranking 304, and gradient boosting ranking 306. Feature overlap analysis 308 also identifies features 2, 4, and 8 as high-ranking features identified by two metrics. In some embodiments, feature overlap analysis 308 may output a subset of features by outputting the features that have been identified as high-ranking by all metrics. In some embodiments, feature overlap analysis 308 may output a subset of features by outputting features that have been identified as high-ranking by one or more metrics. In some embodiments, feature overlap analysis 308 may be graphically represented. In some embodiments, feature overlap analysis 308 may output a list comprising a subset of features.

**[0085]** **FIG. 3B** illustrates an exemplary output 310 of the feature selection process for features used to classify a tumor of a cancer in a subject as HRD-positive or HRD-negative in accordance with some embodiments. Feature importance rankings 312 are shown graphically, and each graph depicts the ranking of features according to a specific feature importance metric. In each graph (ANOVA, random forest, and gradient boosting), each dot represents a feature, with its y-axis value corresponding to its feature importance as calculated by the feature importance metric. In the example of **FIG. 3B**, feature overlap analysis 314 may include the top-ranked features according to each feature importance metric. As shown, the feature overlap analysis can identify the features that are highly ranked by all of the metrics and/or some of the metrics.

[0086] Returning to **FIG. 1**, in some embodiments, the system and method may determine a subset of a plurality of features using an iterative feature selection process 104b in addition to or as an alternative to process 104a. At block 104b, the system evaluates the features using one or more feature importance metrics (e.g., gradient boosting) and then performs an iterative feature selection process to gradually expand a feature set, as described below in **FIG. 4**.

[0087] **FIG. 4** illustrates an iterative feature selection process that may be used by block 104b of **FIG. 1** in accordance with some embodiments. At block 402, the system receives a dataset with a plurality of features (e.g., the plurality of features received at block 102 of **FIG. 1**).

[0088] At block 404 of **FIG. 4**, the system evaluates the features received at block 402 using one or more feature importance metrics (e.g., gradient boosting). The system may then rank the features according to their corresponding feature importance metric scores.

[0089] At block 408 of **FIG. 4**, the system and method obtains a new feature set. In the initial iteration, the system can obtain a new feature set by including the highest-ranking feature(s) as determined by block 404 to the feature set. In a subsequent iteration, the system can expand the existing feature set by adding the next highest-ranking feature(s) as determined by block 404 to obtain a new feature set. The system further obtains a training dataset based with the new feature set. The training dataset can comprise a plurality of data elements, and each data element comprises data related to the new feature set and the corresponding classification label (e.g., HRD-positive or HRD-negative). For example, a data element can comprise data related to the features in the new feature set from a sample and the corresponding classification label (e.g., HRD-positive or HRD-negative) of the sample.

[0090] At block 410 of **FIG. 4**, the system and method trains and evaluates a new classification model using the training dataset from block 408. The system records the model performance in association with the list of features used in the model's training and evaluation. In some embodiments, the training and evaluation of the classification model may be performed using cross-validation methods, as discussed further below by **FIGs. 6A** and **6B**. In some embodiments, the training and evaluation of the classification model may use separate subsets of the dataset from block 408.

[0091] In some embodiments, blocks 408 and 410 of **FIG. 4** are iterated until all the features received in block 402 are included in the data. In each iteration, block 408 adds the next highest-ranked feature(s) to the dataset. For example, in the first iteration, block 408 outputs a feature set comprising the highest-ranking feature and a corresponding training set; in the

second iteration, block 408 outputs a feature set comprising two highest-ranking features and the corresponding training set; in the third iteration, block 408 outputs a feature set comprising three highest-ranking features and the corresponding training set, and so on. In each iteration, block 410 then trains and evaluates a new classification model using the training dataset from block 406. The system iterates blocks 408 and 410 until a condition is met. In some embodiments, the condition comprises block 412, in which the system determines that there are no more features to be added (e.g. all features received at block 402 are included in the dataset used to train and evaluate the classification model at block 410). In some embodiments, the condition comprises a determination that the performance of the new classification model exceeds a threshold. This iterative process allows the system to record the performance of the classification model when trained and evaluated on the highest-ranking feature, the top two highest-ranking features, the top three highest-ranking features, and so on, until all features received at block 402 are used to train a classification model and evaluated for performance. An example of the recorded performance data is shown below in **FIG. 5**.

**[0092]** At block 414 of **FIG. 4**, the system and method utilizes the recorded model performances from block 410 to determine the smallest subset of features that optimizes the performance of the classification model. In some embodiments, the system may determine the smallest subset of features such that adding additional features does not substantially improve model performance. In some embodiments, the system may determine the smallest subset of features such that the classification model performance exceeds a certain predetermined threshold. The subset of features is output at block 414.

**[0093]** **FIG. 5** illustrates an example plot of the model performances determined at block 410 of **FIG. 4**. In the example shown in **FIG. 5**, the horizontal axis indicates the number of high-ranking features included in the data used to train and evaluate the classification models; the vertical axis indicates the performance of the model. In some embodiments, the performance of the model may be evaluated using area under the receiver operating characteristic (ROC) curve (AUC). In the example of **FIG. 5**, it may be determined that the 26 highest-ranking features is output as the subset of features in block 416, although a lower number of features may be selected based on the change in the relative increase in model performance with each added feature.

**[0094]** **FIG. 6A** illustrates an example cross-validation process that may be used to evaluate the performance of a model in accordance with some embodiments. In some embodiments, process 600 may be used at block 410 of **FIG. 4** to evaluate the performance of a model. At

block 602, the system may receive a plurality of data elements. Each of the plurality of data elements may comprise one or more features and a known classification label. At block 604, the system divides the plurality of data elements from block 602 into  $n$  equally-sized subsets. At block 606, the system holds out one of the subsets from block 604 as a “hold-out” set. At block 608, the system trains a model on all data elements that are not held out (e.g. data elements from the  $n-1$  subsets that are not the “hold-out” set). At block 610, the system uses the data elements features from the “hold-out” set as input to the model from block 608. The model generates a plurality of predicted classification labels corresponding to the data elements features. The predicted classification labels are then compared to the known classification labels of the “hold-out” set to evaluate the performance of the model on the “hold-out” set. Blocks 606, 608, and 610 are iterated until all  $n$  subsets from block 604 have been used as the “hold-out” set once. That is, blocks 606, 608, and 610 are iterated  $n$  times, with a different subset used as the “hold-out” set each iteration. Finally, at step 612, the performances from all  $n$  iterations of block 610 are averaged to output an average performance.

**[0095] FIG. 6B** illustrates an example division of the plurality of data elements into five equally sized subsets in accordance with some embodiments. **FIG. 6B** may be an example of **FIG. 6A** where  $n=5$ . A plurality of data elements 622 may be an example of a plurality of data elements from block 602 of **FIG. 6A**. In the example of **FIG. 6B**, plurality of data elements 622 is divided into Set 1, Set 2, Set 3, Set 4, and Set 5. In iteration one 623, at the plurality of data elements 622, Set 1 may be used as the “hold-out” data set as described by block 606. A model may be trained on Set 2, Set 3, Set 4, and Set 5, as described by block 608. The model performance may then be evaluated on “hold-out” data Set 1. This process is then repeated for four more iterations: in iteration two 624, Set 2 is the “hold-out” set, the model is trained on Set 1, Set 3, Set 4, and Set 5, and the model performance is evaluated on Set 2; in iteration three 626, Set 3 is the “hold-out” set, the model is trained on Set 1, Set 2, Set 4, and Set 5, and the model performance is evaluated on Set 3; in iteration four 628, Set 4 is the “hold-out” set, the model is trained on Set 1, Set 2, Set 3, and Set 5, and the model performance is evaluated on Set 4; in iteration five 630, Set 5 is the “hold-out” set, the model is trained on Set 1, Set 2, Set 3, and Set 4, and the model performance is evaluated on Set 5. In the example of **FIG. 6B**, the average performance may be the average of the model performances from iteration one 622, iteration two 624, iteration three 626, iteration four 628, and iteration five 630.

[0096] Returning to **FIG. 1**, at block 106, the system obtains a subset of selected features, as determined by the feature selection of block 104. A classification model 108 is trained using information from selected features 106 and labelled training data 110. In some embodiments, the dataset used for feature selection 104 is the same dataset that is the labelled training data 110. In some embodiments, the dataset used for feature selection 104 is a different dataset from the labelled training data 110. The process of training the classification model is discussed below in the following sections and in **FIG. 7**. Once classification model 108 is trained, features from an unseen tumor of a cancer in a subject (e.g., data elements that are not included in data received in block 102 and are not associated with known classification labels) could be input into model 108 to predict whether the tumor of a cancer in a subject is likely HRD-positive or HRD-negative.

### Data features

[0097] A test sample from a tumor being identified (e.g., classified) can be obtained from a subject. Features, such as basic features, copy number features, and/or short variant features, associated with the test sample include one or more features that can be used as input for the HRD classification model. The HRD classification model is trained based on corresponding features (such as basic features, copy number features, and/or short variant features) from HRD positive data associated with HRD positive samples (such as tumor samples) and HRD negative data associated with HRD negative samples (such as tumor samples). The features can be used as a functional readout of HRD which can help identify tumors with a “BRCAness” profile, which is associated with HRD. Tumors with such HRD-positive phenotypes may be suitable candidates for certain drug therapies that are not (or often not) effective in HRD-negative phenotypes.

[0098] The copy number features can include, but are not limited to, a segment size feature, a number of sequencing reads feature, an absolute copy number feature, a breakpoint count per  $x$  megabases feature, a change point copy number feature, a segment copy number feature, a breakpoint count per chromosome arm feature, and a number of segments with oscillating copy number feature. See Macintyre et al., *Copy-number signatures and mutational processes in ovarian carcinoma*, Nat. Genet. 2018 Sep;50(9):1262-1270. Mixture modeling can be applied to divide each feature distribution into mixtures of Gaussian or mixtures of Poisson distributions to achieve float or binary component features. The copy number features can also include a segment minor allele frequency feature, which is based on the A and B allele frequencies of germline SNPs in the segment.

**[0099]** In some embodiments, the HRD model (e.g., the HRD classifier model) may be trained using more features than used as input. For example, the HRD classification model may be trained based on HRD positive data and HRD negative data each comprising a certain number of features associated with the HRD positive tumors and/or HRD negative tumors. The data input to the HRD classification model may then comprise fewer features. The HRD classifier model may, in one example, adjust a weight for data features omitted from the sample data that is input into the trained HRD classifier model. In addition, the HRD classifier model may be trained using additional data features (such as a measure of genome-wide loss of heterozygosity and/or one or more short variant features, each as described herein), but the data input may, in some embodiments, only comprise one or more copy number features associated with the genome of a tumor associated with a cancer in a subject.

**[0100]** To obtain genomic data features, including copy number features, basic features including measures of gLOH and tumor genome ploidy and/or short variant features, sequencing data is collected by sequencing of at least a portion of at least one genome of a tumor. Absolute or relative copy numbers and segmentation can then be derived from whole genome sequencing data, such as shallow whole genome sequencing (sWGS) data. Circular binary segmentation (CBS) may also be used to partition a genome into segments of constant total copy numbers based on DNA microarray data, from which copy number features may be derived. Alternatively, absolute copy numbers and segmentation can be derived from any technique known in the art, including, but not limited to, exome sequencing (ES) or SNP arrays. The distribution of copy number features can be computed from the absolute copy number data, such as the WGS data. Mixture modeling can be applied to divide each feature distribution into mixtures of Gaussian or mixtures of Poisson distributions to achieve float or binary component features. Thus, a particular “copy number feature” used to train the HRD classification model, or to be inputted into the trained HRD classification model, will be expressed as its component feature. For example, for the copy number feature of segment size, if divided into  $z$  number of components, then there are  $z$  number of possible features which may be used to train the HRD classification model or used to run the HRD classification model. In other words, for a particular test sample, the “copy number feature” in the category of “segment size” (assuming segment size was divided into  $z$  number of components) has  $z$  number of possible inputs, whether for training or running the HRD classification model. If  $z$  is equal to three, then at least one of three segment size features may be input into the HRD classification model: i.e., segsize1, segsize2, or segsize3. Optimal model performance may depend, in part, on the number of component features selected for

each particular category of feature. However, particular categories of features may be divided into any suitable number of component features, and not necessarily those corresponding to a particular probability distribution. Thus, the model may perform well and validate efficiently with more or fewer numbers of component features, even if the performance is not optimal.

**[0101]** When deriving copy number features, the absolute copy number data may first be normalized by matching with a normal dataset to determine the baseline level from which to call copy number variation events. The panel of normal is typically derived from healthy tissue samples (which may be from the same individual from which the tumor is derived from). Analysis of the healthy tissue samples allows for setting a baseline copy number from which to derive the copy number features described herein.

**[0102]** Some of the described copy number features may be assessed across subregions of the genome. For example, a particular copy number feature may be assessed across the centromeric portion of the genome. In another example, a copy number feature may be assessed across the telomeric portion of the genome. In yet a further example, a copy number feature may be assessed across both the telomeric and centromeric portions of the genome. In an exemplary method, to define the telomeric and centromeric portions of the genome, a human reference sequence genome, such as hg19, may be used to define the start and end of each chromosome arm. The length of a particular arm is then divided by two to define the halfway point. For each region analyzed for a copy number feature, a segment falling on the centromeric side of this halfway point is defined as a centromeric segment. A segment falling on the telomeric side of this halfway point is defined as a telomeric segment. If a segment spans the halfway point (for example, a segment beginning on the centromeric side and ending on the telomeric side of the halfway point), then that segment may be designated as both centromeric and telomeric, and may be used in the assessment of both telomeric and centromeric copy number features. Any of the data features described herein, as appropriate, may thus be assessed across the telomeric region of the genome, the centromeric region of the genome, or both the telomeric and centromeric regions of the genome.

**[0103]** Modeling of copy number may be impacted by the estimated base ploidy of the genome being assessed. If the base ploidy is estimated higher, floating-point copy number features may be right-shifted, leading to skewed component scores and ultimately incorrect classifications. Normalizing copy number data to the base ploidy involves dividing copy number data by the mean ploidy of the genome being assessed. Thus, any of the described copy number features may be derived from ploidy-normalized copy number data, wherein the absolute copy numbers are normalized to the mean ploidy of the tumor genome. An example

method to calculate mean ploidy is to take the weighted average copy number for all segments in a sample. For an exemplary method of calculating mean ploidy, *see* Sun et al., *A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal*, PLoS Comput. Biol. 2018 Feb 7;14(2):e1005965.

**[0104]** The features described herein may, in some embodiments, be binned features. Feature binning involves organizing certain values to certain categorical bins. For example, for a feature with values ranging from 0 to 10, a quartile binning may organize each of these values from 0 to 10 into one of four bins, wherein lower values may be organized into a lower bin, and higher values into a higher bin. In some embodiments, the binning is unsupervised. In some embodiments, the binning is supervised. In some embodiments, the binning is equal width binning. In equal width binning, the bins have ranges with approximately the same width. For example, for a feature having values from 1 to 8, equal width binning with four bins would organize values of 1 and 2 into a first bin, values of 3 and 4 to a second bin, and so on. In some embodiments, the binning is equal frequency binning. In equal frequency binning, the bins are organized so that each bin has approximately the same number of values, such that the values are distributed about equally into the bins. For example, for a feature having values from 1 to 10, where lower values are much higher frequency, the binning may organize 1 to a first bin, 2 to a second bin, and 3 to 10 in a third bin. The binning may be binary, tertile, quartile, quintile, sextile, septile, or any other suitable binning organization.

**[0105]** In some embodiments of any the described methods, the copy number features comprise a segment size feature. Segment size is derived from the length in genomic bases of each copy number segment across the genome. For example, if a segment has a copy number of  $x$ , and the next segment has a copy number of  $y$ , then the length of the segment having copy number  $x$  and the length of the segment having copy number  $y$  are factors in the segment size copy number category. In an exemplary embodiment, the distribution of segment size is divided into 10 component features. A lower-numbered segment size feature represents smaller segment sizes (e.g., segsize1), while a higher-numbered segment size feature represent larger segment sizes (e.g., segsize10). In some embodiments, the distribution of segment size is divided into at least 5 component features, such as at least 6, at least 7, at least 8, at least 9, at least 10, or at least 11 component features. In some embodiments, the distribution of segment size is divided into any of 5, 6, 7, 8, 9, 10, or 11 component features. In some embodiments, the segment size feature is assessed across the

telomeric portion of the genome. In some embodiments, the segment size feature is assessed across the centromeric portion of the genome. In some embodiments, the segment size feature is assessed across both the telomeric portion and the centromeric portion of the genome. In some embodiments, the segment size feature is assessed across the entire genome. In some embodiments, the segment size feature is derived from ploidy-normalized copy number data. In some embodiments, the segment size feature is a binned feature.

**[0106]** In some embodiments of any of the described methods, the copy number features comprise a breakpoint count per  $x$  megabases feature. In some embodiments,  $x$  is between about 1 megabases (MB) and about 150 megabases. In some embodiments,  $x$  is any of about 10MB, about 25MB, about 50MB, about 100MB, and about 150MB. Breakpoint count per section represents the number of breakpoints per section across the genome or a portion of the genome. For example, for breakpoint count per 10MB, a processing adjacent window (or, alternatively, a sliding window) of 10MB is analyzed throughout the genome and the number of breakpoints for each frame of the sliding window can then be assessed. It should be noted that although an adjacent window was used in this approach, a sliding window or any other technique suitable for assessing breakpoints count could be used. Regardless, in some exemplary embodiments, breakpoint count per  $x$  megabases is divided into 3 component features. A lower-numbered breakpoint count feature represents fewer breakpoints (e.g., in the case of breakpoint count per 10MB: bp10MB1, indicating fewer breakpoints per frame of a 10MB sliding window or per frame of a 10MB processing adjacent window), while higher-numbered features represent more breakpoints per section (e.g., in the case of breakpoint count per 10MB: bp10MB3, indicating more breakpoints per frame of a 10MB sliding window as compared to a lower-numbered feature, such as bp10MB1). In some embodiments, the distribution of breakpoint count is divided into at least 2 component features, such as at least 3 or at least 4 component features. In some embodiments, breakpoint count per section is divided into any of 2, 3, 4, or 5 component features. In some embodiments, the breakpoint count per  $x$  megabases feature is assessed across the telomeric portion of the genome. In some embodiments, the breakpoint count per  $x$  megabases feature is assessed across the centromeric portion of the genome. In some embodiments, the breakpoint count per  $x$  megabases feature is assessed across the entire genome. In some embodiments, the breakpoint count per  $x$  megabases feature is derived from ploidy-normalized copy number data. In some embodiments, the breakpoint count per  $x$  megabases feature is a binned feature.

**[0107]** In some embodiments of any of the described methods, the copy number features comprise a number of sequencing reads feature obtained from sequencing a genome segment. For a particular genome segment, this value refers to the average number of sequencing reads that align to (i.e., “cover”) the sequenced segment. For genome segments with abnormally high copy number, there will be an increased number of sequencing reads. In contrast, for genome segments that have lost copy number (such as homozygous deletions), there will be fewer sequencing reads. The sequencing reads feature may be expressed as the actual number of reads (such as the average of the reads for each segment analyzed) or a bin of sequencing reads. A lower-numbered sequencing reads feature represents lower absolute sequencing reads, while high-numbered sequencing reads feature represents higher absolute sequencing reads. In some embodiments, sequencing reads feature is assessed across the telomeric portion of the genome. In some embodiments, sequencing reads feature is assessed across the centromeric portion of the genome. In some embodiments, sequencing reads feature is assessed across both the telomeric and centromeric portion of the genome. In some embodiments, sequencing reads feature is derived from ploidy-normalized data. In some embodiments, sequencing reads feature is a binned feature. In some embodiments, the number of sequencing reads feature is a measurement of the number of reads from next generation sequencing (NGS). In some embodiments, the number of sequencing reads feature is expressed as the ratio of sequencing reads for a genome segment in the tumor sample compared to the number of sequencing reads for that genome segment in a control.

**[0108]** In some embodiments of any of the described methods, the copy number features comprise an absolute copy number feature. The absolute copy number may be computed for each genome segment and assigned a value. For example, the assigned values may include 0 (indicating a homozygous deletion), 1 (which may indicate a heterozygous deletion), 2 (which could be a normal count), or more (which may indicate copy number amplification). The absolute copy number feature may represent the actual copy number count (such as the average of the copy number for each segment analyzed) or a bin of copy number values. For example, copy numbers of at least 6 may be binned as representing a high copy number for a segment. Copy numbers between 3 and 5 may be binned as representing a moderately increased copy number. Copy numbers of 1 and 2 may be normal, and copy numbers of 0 may be binned as homozygous deletions. Lower-numbered absolute copy number features represent lower absolute copy number, while high-numbered absolute copy number features represent higher absolute copy number. In some embodiments, absolute copy number is divided into any of 3, 4, 5, 6, 7, 8, or 9 component features. In some embodiments, absolute

copy number feature is assessed across the telomeric portion of the genome. In some embodiments, absolute copy number feature is assessed across the centromeric portion of the genome. In some embodiments, absolute copy number features is assessed across both the telomeric and centromeric portions of the genome. In some embodiments, the absolute copy number feature is derived from ploidy-normalized data. In some embodiments, the absolute copy number feature is a binned feature.

**[0109]** In some embodiments of any of the described methods, the copy number features comprise change point copy number feature. Change point copy number refers to the absolute difference in copy number between genome segments across the genome. For example, adjacent segments modeled at copy numbers of 7 and 2 would have an absolute different of 5. In an exemplary embodiment, the distribution of change point copy number is divided into 7 component features. Lower-numbered change point copy number features represent smaller absolute difference in copy number changes (e.g., changepoint1), while higher-numbered features represent larger absolute difference in copy number changes (e.g., changepoint7). In some embodiments, the distribution of change point copy number is divided into at least 4 component features, such as at least 5, at least 6, at least 7, or at least 8 component features. In some embodiments, change point copy number is divided into any of 4, 5, 6, 7, 8, or 9 component features. In some embodiments, the change point copy number feature is assessed across the telomeric portion of the genome. In some embodiments, the change point copy number feature is assessed across centromeric portion of the genome. In some embodiments, the change point copy number feature is assessed across both the telomeric and centromeric portions of the genome. In some embodiments, the change point copy number feature is derived from ploidy-normalized copy number data. In some embodiments, the change point copy number feature is a binned feature.

**[0110]** In some embodiments of any of the described methods, the copy number features comprise a segment copy number feature. Segment copy number is derived from the copy number of each segment across the genome or a portion of the genome. In an exemplary embodiment, the distribution of segment copy number is divided into 8 component features. Lower-numbered segment copy number features represent lower copy numbers (e.g., copynumber1 may represent a copy number level of 0 or 1, or 0 to 1), while higher-numbered copy number features represent higher copy numbers (e.g., copynumber8). In some embodiments, the distribution of segment copy number is divided into at least 4 component features, such as at least 5, at least 6, at least 7, at least 8, or at least 9 component features. In some embodiments, the distribution of segment copy number is divided into any of 4, 5, 6, 7,

8, 9, or 10 component features. In some embodiments, the segment copy number feature is assessed across the telomeric portion of the genome. In some embodiments, the segment copy number feature is assessed across the centromeric portion of the genome. In some embodiments, the segment copy number feature is assessed across the entire genome. In some embodiments, the segment copy number feature is derived from ploidy-normalized copy number data. In some embodiments, the segment copy number feature is a binned feature.

**[0111]** In some embodiments of any of the described methods, the copy number features comprise a breakpoint count per chromosome arm feature. In an exemplary embodiment, the distribution of breakpoint count per chromosome arm is divided into 5 component features. Lower-numbered breakpoint count per chromosome arm features represents fewer breakpoints per arm (e.g., bpchrarm1), while higher-numbered breakpoint count per chromosome arm features represents more breakpoints per chromosome arm (e.g., bpchrarm5). In some embodiments, the distribution of breakpoint count per chromosome arm is divided into at least 3 component features, such as at least 4, at least 5, at least 6, or at least 7 component features. In some embodiments, the distribution of breakpoint count per chromosome arm is divided into any of 4, 5, 6, 7, or 8 component features. In some embodiments, the breakpoint count per chromosome arm is derived from ploidy-normalized copy number data. In some embodiments, the breakpoint count per chromosome arm feature is a binned feature.

**[0112]** In some embodiments, the copy number features comprise a number of segments with oscillating copy number (osCN) feature. Number of segments with oscillating copy number represents a traversal of the genome or a portion of the genome counting the number of repeated alternating segments between two copy numbers. In an exemplary embodiment, the distribution of number of segments with oscillating copy number is divided into 3 component features. Lower-numbered number of segments with oscillating copy number features represents fewer repeated alternations between two copy numbers (e.g., osCN1), while higher-numbered number of segments with oscillating copy number features represents more repeated alternations between two copy numbers (e.g., osCN3). In some embodiments, the distribution of number of segments with oscillating copy number is divided into at least 2, such as at least 3 or at least 4 component features. In some embodiments, the distribution of number of segments with oscillating copy number is divided into any of 2, 3, 4, or 5 component features. In some embodiments, the number of segments with oscillating copy number feature is assessed across the telomeric portion of the genome. In some embodiments,

the number of segments with oscillating copy number feature is assessed across the centromeric portion of the genome. In some embodiments, the number of segments with oscillating copy number feature is assessed across the entire genome. In some embodiments, the number of segments with oscillating copy number feature is derived from ploidy-normalized copy number data. In some embodiments, the number of segments with oscillating copy number feature is a binned feature.

**[0113]** In some embodiments, the copy number features comprise a segment minor allele frequency (segMAF) feature. The segMAF feature may be derived from either the mean segMAF or the median segMAF of the tumor genome. In a normal genome at a heterozygous allele site, the expected copy number of each allele is 1.0. HRD is associated with the complete loss of an allele (loss of heterozygosity) or an increase in copy number of one allele relative to the other. Thus, segMAF is a traversal of the genome, segment by segment, comparing the ratio of the minor allele to the major allele. Specifically, each heterozygous SNP is analyzed for the A allele and the B allele frequency; the frequency of the minor allele is captured as the minor allele fraction. Balanced loci will have a ratio of about 0.5:0.5 with a minor allele frequency of 0.5. Loss of heterozygosity events will cause an imbalance and skewing of the minor allele frequency to less than about 0.5 for the minor allele fraction. In some embodiments the segMAF feature is assessed across the telomeric portion of the genome. In some embodiments, the segMAF feature is assessed across the centromeric portion of the genome. In some embodiments, the segMAF feature is assessed across the entire genome. In some embodiments, the segment minor allele frequency feature is a binned feature.

**[0114]** The HRD classification model is trained by HRD positive data comprising, for each HRD positive tumor in a plurality of HRD positive tumors, one or more features associated with the HRD positive tumors and a HRD positive label and HRD negative data comprising, for each HRD negative tumor in a plurality of HRD negative training tumors, one or more copy number features associated with the HRD negative tumors and a HRD negative label. The HRD classification model may also be trained based on other features or measures. Accordingly, test data comprising these other features or measures may be inputted into the HRD classification model (including in combination with the one or copy number features). For example, basic features including, for example, a measure of genomic loss of heterozygosity, and/or one or more short variant features, may be used in the HRD classification model (whether to train the HRD classification model or as test data to be inputted to the HRD classification model).

**[0115]** In some embodiments, the basis features comprise an age of the subject from which the tumor was obtained. The patient may be any age, including any of at least 5, at least 10, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, or at least 80 years old. The age feature may be an integer value for the subject. Alternatively, the age feature may be a qualitative feature, such as any of an infant, young, child, young adult, or elderly subject. In some embodiments, the age feature is a binned feature.

**[0116]** In some embodiments, the basic features comprise a cancer type feature. The cancer type feature refers to the tumor origin. The cancer type may include, for example, one of an adrenal, biliary, bone/soft tissue, breast, colon/rectum, esophageal, eye, head and neck, kidney, liver, lung, lymphoid, medulloblastoma, mesothelioma, myeloid, nervous system, neuroendocrine, ovarian, pancreatic, prostate, skin, stomach, testicle, thymus, thyroid, urinary tract, uterine, or vulvar cancer. In some embodiments, the cancer type feature is a binned feature.

**[0117]** In some embodiments, the basic features comprise a cancer stage feature. Staging of cancers is often based on the type of cancer (e.g., pancreatic cancer staging, prostate cancer staging, breast cancer staging, ovarian cancer staging, etc.), although universal staging systems are also known in the art. Any suitable cancer staging system may be used, and may depend, for example, on the location of the tumor, the cell type, the tumor size, the spread and distribution of the tumor, metastasis of the tumor, and the tumor grade. As a data feature, a cancer stage would typically be expressed as ranging from a less severe stage to a higher severity stage. For example, for a cancer stage feature comprising 4 component features, stage1 may indicate an early-stage cancer, while stage4 may indicate a late-stage cancer. In some embodiments, the cancer stage feature is a binned feature.

**[0118]** The HRD positive data and the HRD negative data is typically split into a training dataset, a validation dataset, and/or a testing dataset. During training, the HRD classification model is only provided with the training set. Optionally, the training set may be balanced. Once trained, the model can be validated by performance on the validation set and tuned. The training may be adjusted and repeated in the event the model exhibits over-fitting on the validation set. Once trained, and after optionally validated, the trained model may be evaluated using the testing dataset.

**[0119]** A measure of genomic loss of heterozygosity (gLOH) (e.g., a genome-wide loss of heterozygosity or exome-wide loss of heterozygosity) may be included as a basic feature in some embodiments. The full genome need not be analyzed to determine the genomic loss of

heterozygosity, as whole exome sequencing or targeted sequencing across a large enough portion of the genome may be taken as a proxy from genomic loss of heterozygosity. In some embodiments, the gLOH is encoded as a continuous numeric feature. In some embodiments, the gLOH is encoded as a categorical feature, for example, if the gLOH is above or below a predetermined threshold. The predetermined threshold may be set, for example, at about 10% or higher, about 12% or higher, about 14% or higher, or about 16% or higher. The predetermined threshold may be set, for example, at about 16%. The gLOH may be determined, for example, using the methods described in Swisher et al., *Rucaparib in relapsed, platinum-sensitive high-grade ovarian carcinoma (ARIEL2 Part1): an international, multicenter, open-label, phase 2 trial*, Lancet Oncology, vol. 18, no. 1, pp. 75-87 (2017).

**[0120]** One or more short variant features may be used in the HRD classification model (whether to train the HRD classification model and/or as test data to be inputted to the HRD classification model). These short variant features may include, but are not limited to, one or more of a deletions (such as at least 5-basepair deletion) at, for example, repetitive or microhomology regions feature and/or a mutational signature incorporating two or more short variant features. These short variant features, in an exemplary method, may be identified by comparing the sequencing data corresponding to a tumor sample with a consensus human genome sequence (such as hg19). In some embodiments, the short variant feature is a binned feature.

**[0121]** Multiple short variant features may be combined and expressed as a mutational signature score. For example, the one or more short variant features may comprise a mutational profile, such as one from the COSMIC cancer database. In one example, the one or more short variant features comprise an indel-based signature, such as the COSMIC ID6 or COSMIC ID8 indel signature of the COSMIC cancer database. Sample profiles can be mapped to these COSMIC profiles, for example, using NMF methodology. In another example, the one or more short variant features comprise the COSMIC ID8 of the COSMIC cancer database. In yet another example, the one or more short variant features comprise the SBS3 mutational signature of the COSMIC cancer database. For a summary of exemplary COSMIC ID signatures, see Alexandrov et al., *The repertoire of mutational signatures in human cancer*, Nature 2020; 578(7793):94-101. See also Forbes et al., *COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer*, Nuc. Acids Res. 2011 Jan; 39:D945-D950.

**[0122]** In some embodiments, the one or more short variant features comprise a deletion in microhomology or repetitive regions feature. In some embodiments, the deletions are at least 1-basepair. In some embodiments, the deletions are at least 5-basepairs. Deletions at microhomology regions are a characteristic result of microhomology-mediated end joining (MMEJ), which occurs in the absence of homologous recombination. In this process, short regions of similarity (microhomologies) are used to guide the repair of double stranded breaks in the genome. The identifying characteristic of these deletions is that the 3' end of the deleted sequence will share similarity with the upstream context of the deletion. Thus, the deletions at a microhomology region feature is a measure of the number of deletions that exhibit this behavior and may also be based on the length of the microhomology (i.e., numerous deletions with longer length vs fewer deletions with shorter lengths).

**[0123]** In an exemplary embodiment, the test data comprise a segment minor allele frequency feature and a segment size feature. In some embodiments, the segment minor allele frequency feature is a binned feature. In some embodiments, the segment size feature is a binned feature. The test data may further comprise at least one of a breakpoint count per  $x$  megabases feature, a change point copy number feature, a number of sequencing reads feature, an absolute copy number feature, a segment copy number feature, a breakpoint count per chromosome arm feature, and a number of segment with oscillating copy number feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

**[0124]** In another exemplary embodiment, the test data comprise a segment minor allele frequency feature and a breakpoint count per  $x$  megabases feature. In some embodiments, the segment minor allele frequency feature is a binned feature. In some embodiments, the breakpoint count per  $x$  megabases feature is a binned feature. The test data may further comprise at least one of a segment size feature, a number of sequencing reads feature, an absolute copy number feature, a change point copy number feature, a segment copy number feature, a breakpoint count per chromosome arm feature, and a number of segments with oscillating copy number feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

**[0125]** In another exemplary embodiment, the test data comprise a segment minor allele frequency feature and a change point copy number feature. In some embodiments, the segment minor allele frequency feature is a binned feature. In some embodiments, the change point copy number feature is a binned feature. The test data may further comprise at least one of a segment size feature, a number of sequencing reads feature, an absolute copy number feature, a breakpoint count per  $x$  megabases feature, a segment copy number feature, a breakpoint count per chromosome arm feature, and a number of segments with oscillating copy number feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

**[0126]** In another exemplary embodiment, the test data comprise a segment minor allele frequency feature and a segment copy number feature. In some embodiments, the segment minor allele frequency feature is a binned feature. In some embodiments, the segment copy number feature is a binned feature. The test data may further comprise at least one of a segment size feature, a number of sequencing reads feature, an absolute copy number feature, a breakpoint count per  $x$  megabases feature, a change point copy number feature, a breakpoint count per chromosome arm feature, and a number of segments with oscillating copy number feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

**[0127]** In another exemplary embodiment, the test data comprise a segment minor allele frequency feature and a breakpoint count per chromosome arm feature. In some embodiments, the segment minor allele frequency feature is a binned feature. In some embodiments, the breakpoint count per chromosome arm feature is a binned feature. The test data may further comprise at least one of a segment size feature, a number of sequencing reads feature, an absolute copy number feature, a breakpoint count per  $x$  megabases feature, a change point copy number feature, a segment copy number feature, and a number of segments with oscillating copy number feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

**[0128]** In another exemplary embodiment, the test data comprise a segment minor allele frequency feature and a number of segments with oscillating copy number feature. In some embodiments, the segment minor allele frequency feature is a binned feature. In some embodiments, the number of segments with oscillating copy number feature is a binned feature. The test data may further comprise at least one of a segment size feature, a number of sequencing reads feature, an absolute copy number feature, a breakpoint count per  $x$  megabases feature, a change point copy number feature, a segment copy number feature, and a breakpoint count per chromosome arm feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

**[0129]** In another exemplary embodiment, the test data comprise a segment size feature and a breakpoint count per  $x$  megabases feature. In some embodiments, the segment size feature is a binned feature. In some embodiments, the breakpoint count per  $x$  megabases feature is a binned feature. The test data may further comprise at least one of a segment minor allele frequency (segMAF) feature, a number of sequencing reads feature, an absolute copy number feature, a change point copy number feature, a segment copy number feature, a breakpoint count per chromosome arm feature, and a number of segments with oscillating copy number feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

**[0130]** In another exemplary embodiment, the test data comprise a segment size feature and a change point copy number feature. In some embodiments, the segment size feature is a binned feature. In some embodiments, the change point copy number feature is a binned feature. The test data may further comprise at least one of a segment minor allele frequency (segMAF) feature, a number of sequencing reads feature, an absolute copy number feature, a breakpoint count per  $x$  megabases feature, a segment copy number feature, a breakpoint count per chromosome arm feature, and a number of segments with oscillating copy number feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

**[0131]** In another exemplary embodiment, the test data comprise a segment size feature and a segment copy number feature. In some embodiments, the segment size feature is a binned feature. In some embodiments, the segment copy number is a binned feature. The test data may further comprise at least one of a segment minor allele frequency (segMAF) feature, a number of sequencing reads feature, an absolute copy number feature, a breakpoint count per  $x$  megabases feature, a change point copy number feature, a breakpoint count per chromosome arm feature, and a number of segments with oscillating copy number feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

**[0132]** In another exemplary embodiment, the test data comprise a segment size feature and a breakpoint count per chromosome arm feature. In some embodiments, the segment size feature is a binned feature. In some embodiments, the breakpoint count per chromosome arm feature is a binned feature. The test data may further comprise at least one of a segment minor allele frequency (segMAF) feature, a number of sequencing reads feature, an absolute copy number feature, a breakpoint count per  $x$  megabases feature, a change point copy number feature, a segment copy number feature, and a number of segments with oscillating copy number feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

**[0133]** In another exemplary embodiment, the test data comprise a segment size feature and a number of segments with oscillating copy number feature. In some embodiments, the segment size feature is a binned feature. In some embodiments, the number of segments with oscillating copy number feature is a binned feature. The test data may further comprise at least one of a segment minor allele frequency (segMAF) feature, a number of sequencing reads feature, an absolute copy number feature, a breakpoint count per  $x$  megabases feature, a change point copy number feature, a segment copy number feature, and a breakpoint count per chromosome arm feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

**[0134]** In another exemplary embodiment, the test data comprise a breakpoint count per  $x$  megabases feature and a change point copy number feature. In some embodiments, the breakpoint count per  $x$  megabases feature is a binned feature. In some embodiments, the change point copy number feature is a binned feature. The test data may further comprise at least one of a segment minor allele frequency (segMAF) feature, a number of sequencing reads feature, an absolute copy number feature, a segment size feature, a segment copy number feature, a breakpoint count per chromosome arm feature, and a number of segments with oscillating copy number feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

**[0135]** In another exemplary embodiment, the test data comprise a breakpoint count per  $x$  megabases feature and a segment copy number feature. In some embodiments, the breakpoint count per  $x$  megabases feature is a binned feature. In some embodiments, the segment copy number feature is a binned feature. The test data may further comprise at least one of a segment minor allele frequency (segMAF) feature, a number of sequencing reads feature, an absolute copy number feature, a segment size feature, a change point copy number feature, a breakpoint count per chromosome arm feature, and a number of segments with oscillating copy number feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

**[0136]** In another exemplary embodiment, the test data comprise a breakpoint count per  $x$  megabases feature and a breakpoint count per chromosome arm feature. In some embodiments, the breakpoint count per  $x$  megabases feature is a binned feature. In some embodiments, the breakpoint count per chromosome arm feature is a binned feature. The test data may further comprise at least one of a segment minor allele frequency (segMAF) feature, a number of sequencing reads feature, an absolute copy number feature, a segment size feature, a change point copy number feature, a segment copy number feature, and a number of segments with oscillating copy number feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

[0137] In another exemplary embodiment, the test data comprise a breakpoint count per  $x$  megabases feature and a number of segments with oscillating copy number feature. In some embodiments, the breakpoint count per  $x$  megabases feature is a binned feature. In some embodiments, the number of segments with oscillating copy number feature is a binned feature. The test data may further comprise at least one of a segment minor allele frequency (segMAF) feature, a number of sequencing reads feature, an absolute copy number feature, a segment size feature, a change point copy number feature, a segment copy number feature, and a breakpoint count per chromosome arm feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

[0138] In another exemplary embodiment, the test data comprise a change point copy number feature and a segment copy number feature. In some embodiments, the change point copy number feature is a binned feature. In some embodiments, the segment copy number feature is a binned feature. The test data may further comprise at least one of a segment minor allele frequency (segMAF) feature, a number of sequencing reads feature, an absolute copy number feature, a segment size feature, a breakpoint count per  $x$  megabases feature, a breakpoint count per chromosome arm feature, and a number of segments with oscillating copy number feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

[0139] In another exemplary embodiment, the test data comprise a change point copy number feature and a breakpoint count per chromosome arm feature. In some embodiments, the change point number feature is a binned feature. In some embodiments, the breakpoint count per chromosome arm feature is a binned feature. The test data may further comprise at least one of a segment minor allele frequency (segMAF) feature, a number of sequencing reads feature, an absolute copy number feature, a segment size feature, a breakpoint count per  $x$  megabases feature, a segment copy number feature, and a number of segments with oscillating copy number feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

**[0140]** In another exemplary embodiment, the test data comprise a change point copy number feature and a number of segments with oscillating copy number feature. In some embodiments, the change point copy number feature is a binned feature. In some embodiments, the number of segments with oscillating copy number feature is a binned feature. The test data may further comprise at least one of a segment minor allele frequency (segMAF) feature, a number of sequencing reads feature, an absolute copy number feature, a segment size feature, a breakpoint count per  $x$  megabases feature, a segment copy number feature, and a breakpoint count per chromosome arm feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

**[0141]** In another exemplary embodiment, the test data comprise a segment copy number feature and a breakpoint count per chromosome arm feature. In some embodiments, the segment copy number feature is a binned feature. In some embodiments, the breakpoint count per chromosome arm feature is a binned feature. The test data may further comprise at least one of a segment minor allele frequency (segMAF) feature, a number of sequencing reads feature, an absolute copy number feature, a segment size feature, a breakpoint count per  $x$  megabases feature, a change point copy number feature, and a number of segments with oscillating copy number feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

**[0142]** In another exemplary embodiment, the test data comprise a segment copy number feature and a number of segments with oscillating copy number feature. In some embodiments, the segment copy number feature is a binned feature. In some embodiments, the number of segments with oscillating copy number feature is a binned feature. The test data may further comprise at least one of a segment minor allele frequency (segMAF) feature, a number of sequencing reads feature, an absolute copy number feature, a segment size feature, a breakpoint count per  $x$  megabases feature, a change point copy number feature, and a breakpoint count per chromosome arm feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a

cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

**[0143]** In another exemplary embodiment, the test data comprise a breakpoint count per chromosome arm feature and a number of segments with oscillating copy number feature. In some embodiments, the breakpoint count per chromosome arm feature is a binned feature. In some embodiments, the number of segments with oscillating copy number feature is a binned feature. The test data may further comprise at least one of a segment minor allele frequency (segMAF) feature, a number of sequencing reads feature, an absolute copy number feature, a segment size feature, a breakpoint count per  $x$  megabases feature, a change point copy number feature, and a segment copy number feature. The test data may further comprise a measure of gLOH and/or one or more short variant features. The test data may further comprise one or more of an age of the subject from which the test data were obtained, a cancer type feature, a cancer stage feature, a tumor purity feature, and a tumor genome ploidy feature.

### **HRD model**

**[0144]** A tumor of a cancer in a subject is classified using a trained HRD classification model that is configured to classify the tumor as HRD-positive (or likely HRD positive) or HRD-negative (or likely HRD negative). The HRD classification model is trained using HRD positive data comprising, for each HRD-positive tumor in a plurality of HRD-positive tumors, one or more data features (such as one or more copy number features and/or one or more short variant features, among other possible features) associated with the HRD-positive tumors and a HRD-positive label. The HRD classification model is further trained using HRD negative data comprising, for each HRD-negative tumor in a plurality of HRD-negative tumors, one or more data features (such as one or more copy number features and/or one or more short variant features, among other possible features) associated with the HRD-negative tumors and a HRD-negative label. Test data comprising one or more data features (such as one or more copy number features and/or one or more short variant features, among other possible features) associated with a genome of a tumor in a subject is input into the trained HRD classification model, which then classifies the tumor as HRD-positive (or likely HRD positive) or HRD-negative (or likely HRD negative) based on the test data.

**[0145]** The models described herein can include one or more machine-learning models, one or more non-machine-learning models, or any combination thereof. The machine-learning models described herein include any computer algorithms that improve automatically through

experience and by the use of data. The machine-learning models can include supervised models, unsupervised models, semi-supervised models, self-supervised models, etc.

Exemplary machine-learning models include, but are not limited to: linear regression, logistic regression, decision tree, SVM, naïve Bayes, neural networks, K-Means, analysis of variance (ANOVA), Chi-Square analysis, random forest, dimensionality reduction algorithms, and gradient boosting algorithms (such as XGB). The non-machine-learning models can include any computer algorithms that do not necessarily require training and retraining.

**[0146]** The HRD classifier may be a probabilistic classifier, such as a gradient boosting model. The probabilistic classifier can be configured to compute a probability that the tumor is HRD positive or HRD negative, such as by outputting a HRD positive likelihood score or a HRD negative likelihood score. Based on the probability or probabilities outputted from the HRD classification model, the tumor can be called as being HRD positive or HRD negative. Optionally, the tumor may be called as ambiguous, for example if neither the probability that the tumor is HRD positive nor that the probability that the tumor is HRD negative is above a predetermined probability threshold. The HRD positive data and the HRD negative data can include the copy number features and/or the short variant features described herein.

**[0147]** The HRD negative data may comprise genomes with wild-type alleles (i.e., alleles not associated with HRD) at certain HRD-associated genes. For example, in some embodiments, the HRD negative data comprises data associated with genomes with wild-type alleles at one or more of a gene associated with HRD, including, but not limited to, *BRCA1*, *BRCA2*, *ATM*, *BARD1*, *BRIP1*, *CDK12*, *CHEK1*, *CHEK2*, *FANCL*, *PALB2*, *RAD51B*, *RAD51C*, *RAD51D*, and/or *RAD45L*. In some embodiments, the HRD negative data comprises promoter methylation data of one or more of a gene associated with HRD, including, but not limited to, *BRCA1*, *BRCA2*, *ATM*, *BARD1*, *BRIP1*, *CDK12*, *CHEK1*, *CHEK2*, *FANCL*, *PALB2*, *RAD51B*, *RAD51C*, *RAD51D*, and/or *RAD45L*. In some embodiments, the HRD negative data comprises RNA expression data of one or more of a gene associated with HRD, including, but not limited to, *BRCA1*, *BRCA2*, *ATM*, *BARD1*, *BRIP1*, *CDK12*, *CHEK1*, *CHEK2*, *FANCL*, *PALB2*, *RAD51B*, *RAD51C*, *RAD51D*, and/or *RAD45L*. In some embodiments, the HRD negative data comprises data associated with genomes associated with tumors that were found to be resistant to platinum-based drugs (e.g., chemotherapy) and/or PARP inhibitors. In some embodiments, the HRD negative data comprises data associated with genomes associated with tumors previously classified as HRD negative. In some embodiments, the HRD negative data is, at least in part, derived from a consensus human genome sequence, or a portion thereof.

[0148] The HRD positive data may comprise data associated with genomes with HRD-associated alleles at certain HRD-associated genes. For example, in some embodiments, the HRD positive data comprises data associated with genomes with mutations at one or more of a gene associated with HRD, including, but not limited to, *BRCA1*, *BRCA2*, *ATM*, *BARD1*, *BRIP1*, *CDK12*, *CHEK1*, *CHEK2*, *FANCL*, *PALB2*, *RAD51B*, *RAD51C*, *RAD51D*, and/or *RAD45L*, particularly biallelic mutations thereof. In some embodiments, the HRD positive data comprises promoter methylation data of one or more of a gene associated with HRD, including, but not limited to, *BRCA1*, *BRCA2*, *ATM*, *BARD1*, *BRIP1*, *CDK12*, *CHEK1*, *CHEK2*, *FANCL*, *PALB2*, *RAD51B*, *RAD51C*, *RAD51D*, and/or *RAD45L*. In some embodiments, the HRD positive data comprises RNA expression data of one or more of a gene associated with HRD, including, but not limited to, *BRCA1*, *BRCA2*, *ATM*, *BARD1*, *BRIP1*, *CDK12*, *CHEK1*, *CHEK2*, *FANCL*, *PALB2*, *RAD51B*, *RAD51C*, *RAD51D*, and/or *RAD45L*. In some embodiments, the HRD positive data comprises data associated with genomes associated with tumors that were found to be sensitive to platinum-based drugs and/or PARP inhibitors. In some embodiments, the HRD positive data comprises data associated with genomes associated with tumors previously classified as HRD positive. In some embodiments, the HRD positive data comprises data associated with tumors having biallelic *BRCA1* and *BRCA2* mutations associated with HRD.

[0149] The HRD positive data may be balanced with the HRD negative data. For example, in an unbalanced training dataset, the number of HRD positive training tumors may outnumber the number of HRD negative tumors (or vice versa). Balancing the data ensures the model has a sufficient number of each label to avoid biasing to one label. When balanced, the number of HRD positive tumors or the number of HRD negative tumors are adjusted so that the ratio between them is at a desired level (such as approximately 1:1 or any other desired ratio). Using the balanced dataset, the HRD classifier may be trained and then tested against a test dataset comprising HRD positive tumors and HRD negative tumors.

[0150] The tumors used to train the HRD classifier each comprise an HRD positive label or a HRD negative label. Any suitable methodology may be used to computationally label (e.g., apply a metadata tag to) the tumors as HRD positive or HRD negative. An HRD positive label may be assigned by the presence of alterations in one of the HRD-associated genes, such as one of a gene associated with HRD, including, but not limited to, *BRCA1*, *BRCA2*, *ATM*, *BARD1*, *BRIP1*, *CDK12*, *CHEK1*, *CHEK2*, *FANCL*, *PALB2*, *RAD51B*, *RAD51C*, *RAD51D*, and/or *RAD45L*, particularly biallelic alterations thereof. Mutations in one or both of *BRCA1* and *BRCA2* are especially indicative of HRD positivity, especially biallelic

*BRCA1/BRCA2* mutations. Tumors may also be labeled as HRD positive based on clinical history. For example, if a tumor was sensitive to a PARP inhibitor or a platinum-based drug regimen, then the tumor is more likely to be HRD positive. An HRD negative label may be assigned based on the absence of alterations in one of the HRD-associated genes, such as one of a gene associated with HRD, including, but not limited to, *BRCA1*, *BRCA2*, *ATM*, *BARD1*, *BRIP1*, *CDK12*, *CHEK1*, *CHEK2*, *FANCL*, *PALB2*, *RAD51B*, *RAD51C*, *RAD51D*, and/or *RAD45L*. Mutations in HRD-associated genes may be detected by comparison of the gene sequence with a reference genome, such as a consensus human genome sequence such as hg19. Likewise, tumors may also be labeled as HRD negative based on clinical history. For example, if a tumor was resistant to a PARP inhibitor or a platinum-based drug regimen, then the tumor is more likely to be HRD negative. This is especially true if the tumor was treatment naïve prior to treatment with the PARP inhibitor or platinum-based drug regimen, since HRD positive tumors may develop resistance to these drugs after rounds of treatment. Although each tumor may comprise an HRD positive or HRD negative label, this label does not require absolute certainty that a tumor is HRD positive or HRD negative. Instead, given a robust training dataset comprising numerous HRD positive tumors and numerous HRD negative tumors, and by avoiding overfitting of these data as is known in the art, the contributions of false positives and false negatives are averaged out in the model. Further, the use of a larger training dataset, particularly a balanced training dataset and a dataset having well-defined positive and negative labels (such as by using validated consensus genomes for HRD-negative labels; and by using validated biallelic *BRCA1/2* mutants or validated, well-characterized *BRCAness* samples for HRD-positive labels), allows the model to properly assess the nuanced differences between HRD-negative phenotypes and those exhibiting HRD scarring (i.e., HRD-positive phenotypes).

**[0151]** The classification method is a computer-implemented method. This classification may be executed on a specifically configured machine or system that includes program instructions for executing a trained HRD classifier model, which may be stored on a non-transitory computer readable memory of the computer or system. The computer generally includes one or more processors that can access the memory. The one or more processors can receive data (e.g., test data such as one or more copy number features and/or one or more short variant features associated with a genome of a tumor in a subject and, in some embodiments, other features and measures), which may also be stored on the memory. The one or more processors can access the trained HRD classifier model, and can input the test

data into the model. The one or more processors and the trained HRD classifier model can then classify the cancer as likely HRD positive or likely HRD negative.

**[0152]** The HRD classifier model may classify the tumor of the cancer as HRD positive or HRD negative. In some embodiments, the HRD classifier model may classify the tumor as likely HRD positive, likely HRD negative, or ambiguous. For example, the HRD classifier model may classify the tumor as ambiguous if it cannot classify the tumor as likely HRD positive or likely HRD negative with sufficiently high confidence or probability. The confidence or probability threshold may be set by the user as desired, given the tolerance for inaccurate classification. In one example, the user may set the HRD-positive likelihood score threshold at 0.8 and the HRD-negative likelihood score threshold at 0.2. If the HRD-positive likelihood score is below 0.8 and/or if the HRD-negative likelihood score is above 0.2, then the HRD model may not classify the tumor as HRD positive, and would either classify the tumor as HRD negative (depending on how low the HRD-positive likelihood score is and how high the HRD-negative likelihood score is) or ambiguous.

**[0153]** In some embodiments, the HRD classifier outputs a likelihood score that the tumor is HRD positive. In some embodiments, the HRD classifier outputs a likelihood score that the tumor is HRD negative. The HRD classifier may be configured to output either or both of an HRD positive likelihood score and an HRD negative likelihood score. The HRD classifier may also be configured to output a ratio of the HRD positive likelihood score to the HRD negative likelihood score and/or a ratio of the HRD negative likelihood score to the HRD positive likelihood score. The likelihood scores may be expressed as a value from 0.0 (indicating a certainty that the tumor is not HRD positive or HRD negative) to 1.0 (indicating a certainty that the tumor is HRD positive or HRD negative). For example, the trained HRD classifier may receive test sample data comprising a plurality of data features associated with a tumor of a cancer in a subject and output an HRD positive likelihood score of 0.8 and an HRD negative likelihood score of 0.15. The HRD classifier may be configured to call the tumor as HRD positive or HRD negative based upon the likelihood score or scores. In the preceding example, based on the HRD positive likelihood score 0.8 and the HRD negative likelihood score of 0.15, the HRD classifier may call the tumor as HRD positive. In some embodiments, the HRD classifier will call the tumor as HRD positive if the HRD positive likelihood score is at least 0.4, such as at least 0.45, at least 0.5, at least 0.55, at least 0.6, at least 0.65, at least 0.70, at least 0.75, at least 0.80, at least 0.85, at least 0.90, at least 0.95, or at least 0.99. In some embodiments, the HRD classifier will call the tumor as HRD positive if the HRD positive likelihood score is at least 0.7. In some embodiments, the HRD classifier

will call the tumor as HRD positive if the HRD positive likelihood score is at least 0.8. In some embodiments, the HRD classifier will call the tumor as HRD positive if the HRD positive likelihood score is at least 0.9. In some embodiments, the HRD classifier will call the tumor as HRD negative if the HRD negative likelihood score is at least 0.4, such as at least 0.5, at least 0.6, at least 0.65, at least 0.70, at least 0.75, at least 0.80, at least 0.85, at least 0.90, at least 0.95, or at least 0.99. In some embodiments, the HRD classifier will call the tumor as HRD negative if the HRD negative likelihood score is at least 0.7. In some embodiments, the HRD classifier will call the tumor as HRD negative if the HRD negative likelihood score is at least 0.8. In some embodiments, the HRD classifier will call the tumor as HRD negative if the HRD negative likelihood score is at least 0.9. In some embodiments, the HRD classifier will call the tumor as HRD positive if the HRD negative likelihood score is less than 0.5, such as less than 0.45, less than 0.40, less than 0.35, less than 0.30, less than 0.30, less than 0.25, less than 0.20, less than 0.15, less than 0.10, or less than 0.05. In some embodiments, the HRD classifier will call the tumor as HRD negative if the HRD positive likelihood score is less than 0.5, such as less than 0.45, less than 0.40, less than 0.35, less than 0.30, less than 0.30, less than 0.25, less than 0.20, less than 0.15, less than 0.10, or less than 0.05. In some embodiments, the HRD classifier will call the tumor as HRD positive if the HRD positive likelihood score is above a certain threshold (such as at least 0.80) and the HRD negative likelihood score is below a certain threshold (such as less than 0.25). In some embodiments, the HRD classifier will call the tumor as HRD negative if the HRD negative likelihood score is above a certain threshold (such as at least 0.80) and the HRD positive likelihood score is below a certain threshold (such as less than 0.25). In some embodiments, the HRD classifier will call the tumor as ambiguous if the HRD positive likelihood score is below a certain threshold and the HRD negative likelihood score is below threshold, or if the absolute values of the likelihood scores are within a threshold percent similarity.

**[0154]** A report may be generated that identifies the cancer as likely HRD positive or likely HRD negative (or ambiguous). The report may be, for example, an electronic medical record or a printed report, which can be transmitted to the subject or a healthcare provider (such as a doctor, a nurse, a clinic, etc.) associated with the subject. The report may be used to make healthcare decisions, such as the method or drug by which the tumor of the cancer is treated.

**[0155]** The report may be displayed on an electronic display or customized interface. For example, in some embodiments, the computer-implemented method may automatically generate the report, and may automatically display the generated report on an electronic display or customized interface.

[0156] FIG. 7 shows an exemplary method for training and operating the HRD classification model 702 configured to classify a tumor of a cancer in a subject as HRD-positive or HRD-negative. The HRD classification model 702 is trained using a data set comprising an HRD positive training data set 704 and an HRD negative training data set 706. The HRD positive training dataset 704 includes one or more HRD positive sample data elements (i.e., HRD positive sample 1 data through HRD positive sample *i*). Each HRD positive sample data element is associated with features (e.g., copy number features, basic features, short variant features, etc.) for HRD positive tumors. The HRD positive sample data element may also include other data features, such as a measure of gLOH and/or short variant features (not shown). The features are labeled as being associated with a HRD positive label. Similarly, the HRD negative training dataset 706 includes one or more HRD negative training sample data elements (i.e., HRD(-) sample 1 through HRD(-) sample *j*). Each HRD negative sample data element is associated with features (e.g., copy number features, basic features, short variant features, etc.) for HRD negative tumors. The HRD negative sample data element may also include other data features, such as a measure of gLOH and/or short variant features (not shown). The HRD negative samples are labeled as being associated with HRD negative label.

[0157] In some embodiments, the HRD classification model 702 is a tree-based gradient boosting model (such as XGBoost). In this model, rather than training all of the models in isolation of one another (e.g., by a random forest), the model is trained in succession such that each new model fits the residuals from the previous models. Therefore, the model achieves a strong classifier from many sequentially-connected weaker classifiers. Repeated cross-validation may be used in the training data for estimating the performance of the HRD classification models.

[0158] After classification model 702 has been trained on the training dataset, the classification model 702 may be used to classify a tumor of a cancer in a subject as HRD-positive or HRD-negative. To classify a tumor of a cancer in a subject as HRD-positive or HRD-negative, classification model 702 receives test data 708 comprising test feature data associated with the tumor to be classified. The test data 708 includes one or more copy number features and may include one or more basic features, one or more short variant features, etc. The classification model 702 may determine a probability that the tumor is HRD positive 710 and/or a probability that the tumor is HRD negative 712. The probabilities 710 and 712 are optionally inputted into a HRD calling module 714. The HRD calling module 714 can call the cancer as HRD positive or HRD negative. For example, if the probability that the tumor test sample is HRD positive 710 is greater than the probability that the tumor test

sample is HRD negative 712, then the tumor test sample can be called as HRD positive. If the probability that the tumor test sample is HRD negative 712 is greater than the probability that the tumor test sample is HRD positive 710, then the tumor test sample can be called as HRD negative. Optionally, if neither of the probabilities 710 and 712 are above a predetermined threshold, the tumor test sample can be called as ambiguous.

**[0159]** The methods described herein may be implemented using one or more computer systems. Such computer systems can include one or more programs configured to execute one or more processors for the computer system to perform such methods. One or more steps of the computer-implemented methods may be performed automatically. The computer system may include one or more computing nodes. For example, a system may include two or more computing nodes (e.g., servers, computers, routers, or other types of electronic devices that include a network interface), which may be connected and configured to communicate and execute the methods over said network on one or more computing nodes of the network.

**[0160]** **FIG. 8** shows an example of a computing device in accordance with one embodiment. Device 1100 can be a host computer connected to a network. Device 1100 can be a client computer or a server. As shown in **FIG. 8**, device 1100 can be any suitable type of microprocessor-based device, such as a personal computer, workstation, server or handheld computing device (portable electronic device) such as a phone or tablet. The device can include, for example, one or more of processor 1110, input device 1120, output device 1130, storage 1140, and communication device 1160. Input device 1120 and output device 1130 can generally correspond to those described above, and can either be connectable or integrated with the computer.

**[0161]** Input device 1120 can be any suitable device that provides input, such as a touch screen, keyboard or keypad, mouse, or voice-recognition device. Output device 1130 can be any suitable device that provides output, such as a display, touch screen, haptics device, or speaker.

**[0162]** Storage 1140 can be any suitable device that provides storage, such as an electrical, magnetic or optical memory including RAM, cache, hard drive, or removable storage disk. Communication device 1160 can include any suitable device capable of transmitting and receiving signals over a network, such as a network interface chip or device. The components of the computer can be connected in any suitable manner, such as via a physical bus or wirelessly.

**[0163]** The HRD Classification Module 1150, which can be stored in storage 1140 and executed by processor 1110, can include, for example, one or more program instructions for executing and implementing the methods and process associated with the HRD model (e.g., as embodied in the devices as described above).

**[0164]** The HRD Classification Module 1150 can also be stored and/or transported within any non-transitory computer-readable storage medium for use by or in connection with an instruction execution system, apparatus, or device, such as those described above, that can fetch instructions associated with the software from the instruction execution system, apparatus, or device and execute the instructions. In the context of this disclosure, a computer-readable storage medium can be any medium, such as storage 1140, that can contain or store programming for use by or in connection with an instruction execution system, apparatus, or device.

**[0165]** The HRD Classification Module 1150 can also be propagated within any transport medium for use by or in connection with an instruction execution system, apparatus, or device, such as those described above, that can fetch instructions associated with the software from the instruction execution system, apparatus, or device and execute the instructions. In the context of this disclosure, a transport medium can be any medium that can communicate, propagate or transport programming for use by or in connection with an instruction execution system, apparatus, or device. The transport readable medium can include, but is not limited to, an electronic, magnetic, optical, electromagnetic or infrared wired or wireless propagation medium.

**[0166]** Device 1100 may be connected to a network, which can be any suitable type of interconnected communication system. The network can implement any suitable communications protocol and can be secured by any suitable security protocol. The network can comprise network links of any suitable arrangement that can implement the transmission and reception of network signals, such as wireless network connections, T1 or T3 lines, cable networks, DSL, or telephone lines.

**[0167]** Device 1100 can implement any operating system suitable for operating on the network. Software 350 can be written in any suitable programming language, such as C, C++, Java or Python. In various embodiments, application software embodying the functionality of the present disclosure can be deployed in different configurations, such as in a client/server arrangement or through a Web browser as a Web-based application or Web service, for example.

**Treatment methods**

**[0168]** Characterization of a tumor as HRD-positive or HRD-negative (or likely HRD-positive or likely HRD-negative) is particularly useful for selecting an effective treatment for a subject having the tumor. Tumors classified as HRD-positive are often more sensitive to certain drugs and therapies that HRD-negative tumors may be resistant to. Based on the classification of a tumor as HRD-positive, likely HRD-positive, HRD-negative, or likely HRD-negative, different drugs or therapies may be selected. Thus, a method of treating cancer in a subject can include assessing a tumor of the cancer as likely HRD positive or likely HRD negative (or calling a tumor of the cancer as HRD positive or HRD negative) according to the methods described herein and then administering to the subject a therapeutically effective amount of a drug based on the classification of the tumor as likely HRD positive or likely HRD negative (or based on the call of the tumor as HRD positive or HRD negative).

**[0169]** The method of treating a cancer in a subject can include obtaining a classification of a tumor of the cancer in the subject as likely HRD positive or likely HRD negative. To obtain this classification, the HRD classification model described herein may be used. One or more copy number features associated with a genome of the tumor of the cancer may be inputted into the HRD classification model which is configured to classify the tumor, based on the one or more copy number features associated with the genome of the tumor in the subject, as likely HRD positive or likely HRD negative. The HRD classification model is trained using HRD positive data from a plurality of HRD positive tumors and HRD negative data from a plurality of HRD negative tumors. The classification may be obtained, for example, by operating the HRD classification model, or by receiving the results from another that operated the HRD classification model.

**[0170]** One or more basic features and/or one or more short variant features may be inputted into the HRD classification model which is configured to classify the tumor based on the one or more basic features and/or the one or more short variant features, as likely HRD positive or likely HRD negative. The one or more short variant features and the one or more basic features may be in addition to, or in the alternative to, the one or more copy number features.

**[0171]** In some embodiments, the treatment methods may include obtaining the test sample data, including the one or more copy number features. In some embodiments, the treatment methods may comprise obtaining the one or more basic features. In some embodiments, the treatment methods may include obtaining the measure of genome-wide loss of heterozygosity. In some embodiments, the treatment methods may include obtaining the one

or more short variant features. A test sample may be obtained from the subject, and nucleic acid molecules may be derived from the test sample. The test sample may be, for example, a solid tissue biopsy of the cancer, and nucleic acids may be isolated from the solid tissue sample. Optionally, the test sample may be preserved, for example, by freezing the test sample or fixing the sample (e.g., by forming a formalin-fixed paraffin-embedded (FFPE) sample) prior to isolating the nucleic acid molecules. Alternatively, the test sample is a liquid biopsy sample (e.g., a blood, plasma, or other liquid sample from the subject), and nucleic acids, including circulating tumor DNA (ctDNA), may be obtained from the liquid sample. The nucleic acids from the sample may be assayed and then analyzed to generate any of the one or more copy number features, the one or more basic features, or the one or more short variant features.

**[0172]** Obtaining the classification of the tumor as likely HRD positive or likely HRD negative can include inputting the described features and/or measures into the HRD classification model and classifying, using the features and/or measures, the cancer as likely HRD positive or likely HRD negative based on the data input to the HRD classification model. Alternatively, obtaining the classification of the tumor as likely HRD positive or likely HRD negative may include receiving a report from another entity. The report may be generated by the other entity, and the report can include a classification of the tumor as likely HRD positive or likely HRD negative, wherein the classification is generated using the HRD classification model described herein. In some embodiments, the report includes a likelihood score that the tumor is HRD positive and/or a likelihood score that the tumor is HRD negative, and a final classification can be made based on the likelihood score(s).

**[0173]** Once a classification of the tumor as likely HRD positive or likely HRD negative has been made, a treatment can be selected based on the classification. If the tumor is classified as likely HRD positive, a treatment that is effective in a HRD positive tumor is selected. The selected treatment can then be administered to the subject to treat the tumor that is classified as likely HRD positive. If the tumor is classified as likely HRD negative, a treatment that is not a platinum-based drug or a PARP inhibitor may be selected. The selected treatment can then be administered to the subject to treat the tumor that is classified as likely HRD negative.

**[0174]** Treatments that are effective in a HRD positive tumor can include one or more PARP inhibitors and/or one or more platinum-based agents. PARP inhibitors may include, but are not limited to, veliparib, olaparib, talazoparib, iniparib, rucaparib, and niraparib. PARP inhibitors are described in Murphy and Muggia, *PARP inhibitors: clinical development*,

*emerging differences, and the current therapeutic issues*, Cancer Drug Resist 2019;2:665-79.

Platinum-based agents may include, but are not limited to, cisplatin, oxaliplatin, and carboplatin. Platinum-based drugs are described in Rottenberg et al., *The rediscovery of platinum-based cancer therapy*, Nat. Rev. Cancer 2021 Jan;21(1):37-50.

**[0175]** The tumor to be treated is a tumor in a subject. In one embodiment, the tumor is a pancreatic cancer. In another embodiment, the tumor is a prostate cancer. In some embodiments, the tumor is an ovarian, breast, or prostate cancer. In some embodiments, the tumor is a tumor associated with HRD, which may include, but is not limited to, one of adrenal, biliary, bone/soft tissue, breast, colon/rectum, esophageal, eye, head and neck, kidney, liver, lung, lymphoid, medulloblastoma, mesothelioma, myeloid, nervous system, neuroendocrine, ovarian, pancreatic, prostate, skin, stomach, testicle, thymus, thyroid, urinary tract, uterine, or vulvar cancer. *See* Nguyen et al., *Pan-cancer landscape of homologous recombination deficiency*, Nat. Commun. 2020 Nov 4;11(1):5584.

**[0176]** Although the disclosure has been fully described with reference to the accompanying figures, it is to be noted that various changes and modifications will become apparent to those skilled in the art. Such changes and modifications are to be understood as being included within the scope of the disclosure as defined by the claims.

**[0177]** The foregoing description, for purpose of explanation, has been described with reference to specific embodiments. However, the illustrative discussions above are not intended to be exhaustive or to limit the invention to the precise forms disclosed. Many modifications and variations are possible in view of the above teachings. The embodiments were chosen and described in order to best explain the principles of the techniques and their practical applications. Others skilled in the art are thereby enabled to best utilize the techniques and various embodiments with various modifications as are suited to the particular use contemplated.

## CLAIMS

In the claims

1. A method, comprising:
  - providing a genome obtained from a tumor of a subject;
  - optionally, ligating one or more adapters onto the genome;
  - amplifying nucleic acid molecules from the genome;
  - capturing nucleic acid molecules from the amplified genome, wherein the captured nucleic acid molecules are captured by hybridization to one or more bait molecules;
  - deriving, from the captured nucleic acid molecules, a set of input features;
  - inputting, by one or more processors, the set of input features into a trained homologous recombination deficiency (HRD) model to identify the tumor as HRD-positive or HRD-negative using the trained HRD model, wherein the model is trained by:
    - determining one or more feature importance metrics associated with each feature of a plurality of features,
    - identifying a subset of features in the plurality of features using the one or more feature importance metrics, and
    - training, by the one or more processors, the HRD model based on the identified subset of features; and
  - classifying, by the one or more processors, using the trained HRD model, the tumor as HRD-positive or HRD-negative.
2. A method, comprising:
  - receiving, by one or more processors, a plurality of features;
  - identifying, by the one or more processors, a subset of features in the plurality of features using one or more feature importance metrics; and
  - training, by the one or more processors, a homologous recombination deficiency (HRD) model based on the identified subset of the plurality of features,
    - wherein the HRD model is configured to receive sample data associated with a genome of a tumor in a subject and identify the tumor in the subject as HRD-positive or HRD-negative using the sample data.
3. A method, comprising:

receiving, by one or more processors, sample data associated with a genome of a tumor in a subject;

inputting, by the one or more processors, the sample data into a trained homologous recombination deficiency (HRD) model, wherein the HRD model is trained by:

determining one or more feature importance metrics associated with each feature of a plurality of features,

identifying a subset of features in the plurality of features using the one or more feature importance metrics, and

training, by the one or more processors, the HRD model based on the identified subset of features; and

classifying, by the one or more processors, using the trained HRD model, the tumor as HRD-positive or HRD-negative.

4. The method of any one of claims 1-3, wherein the plurality of features comprises one or more copy number features, one or more short variant features, or a combination thereof.

5. The method of any one of claims 1-4, wherein the one or more feature importance metrics comprise one or more of a Chi-Square test, analysis of variance (ANOVA), random forest, or gradient boosting.

6. The method of any one of claims 1-5, wherein identifying the subset of features in the plurality of features comprises:

obtaining, by the one or more processors, one or more feature rankings according to the one or more feature importance metrics; and

selecting, by the one or more processors, the subset of the plurality of features based on one or more feature rankings.

7. The method of any one of claims 1-5, wherein identifying the subset of the plurality of features comprises:

(a) obtaining, by one or more processors, a feature ranking of the plurality of features according to a feature importance metric;

(b) obtaining, by the one or more processors, a new feature set by adding one or more features from the plurality of features to an existing feature set based on the feature ranking;

- (c) training, by the one or more processors, a new HRD model using the new feature set;
  - (d) evaluating, by the one or more processors, the trained new HRD model to obtain an evaluation result; and
  - (e) storing, by the one or more processors, the evaluation result associated with the new HRD model and the new feature set;
  - (f) repeating, by the one or more processors, steps (b)-(e) to obtain a plurality of evaluation results until a condition is met; and
  - (g) selecting, by the one or more processors, the subset of the plurality of features based on the plurality of evaluation results.
8. The method of any one of claims 1-7, wherein the trained HRD model is a classification model, the method further comprising:
- receiving new sample data associated with a genome of a tumor in a new subject, wherein the new sample data is related to the subset of the plurality of features;
  - providing the new sample data to the trained HRD classification model to produce a classification result of HRD-positive or HRD-negative; and
  - outputting the classification result.
9. The method of claim 8, wherein the classification result comprises at least one of a HRD-positive likelihood score and a HRD-negative likelihood score.
10. The method of any one of claims 1-9, wherein the HRD model is a classification model, a regression model, a neural network, or any combination thereof.
11. The method of claim 9 or claim 10, comprising recording, in a digital electronic file associated with the new subject, at least one of the HRD-positive likelihood score and the HRD-negative likelihood score.
12. The method of any one of claims 9-11, comprising recording in a digital electronic file associated with the new subject that the tumor is HRD positive based on the HRD positive likelihood score or a designation that the tumor is HRD negative based on the HRD negative likelihood score.

13. The method of any one of claims 1-12, wherein the plurality of features comprise at least one of a segment minor allele frequency (segMAF) feature, a number of sequencing reads feature, a segment size feature, a breakpoint count per  $x$  megabases feature, a change point copy number feature, a segment copy number feature, a breakpoint count per chromosome arm feature, or a number of segments with oscillating copy number feature.
14. The method of any one of claims 1-13, wherein at least one of the plurality of features is assessed across the centromeric portion of the genome.
15. The method of any one of claims 1-14, wherein at least one of the plurality of features is assessed across the telomeric portion of the genome.
16. The method of any one of claims 1-15, wherein at least one of the plurality of features is assessed across both the centromeric and telomeric portions of the genome.
17. The method of any one of claims 1-16, wherein the plurality of features comprise a breakpoint count per  $x$  megabases feature, wherein the breakpoint count per  $x$  megabases feature is based on the number of breakpoints appearing in windows of  $x$  megabases in length across the genome.
18. The method of claim 17, wherein breakpoint count per  $x$  megabases feature is assessed across: (i) the telomeric portion of the genome; (ii) the centromeric portion of the genome; or (iii) both the telomeric portion and the centromeric portion of the genome.
19. The method of claim 17 or claim 18, wherein  $x$  is between about 1 and about 100 megabases.
20. The method of any one of claims 17-19, wherein  $x$  is about 10 megabases, about 25 megabases, about 50 megabases, or about 100 megabases.
21. The method of any one of claims 17-20, wherein the breakpoint count per  $x$  megabases feature is a binned feature.

22. The method of any one of claims 1-21, wherein the plurality of features comprise a change point copy number feature, wherein the change point copy number is based on the absolute difference in copy number between adjacent genome segments across the genome of the tumor of the subject.
23. The method of claim 22, wherein the change point copy number feature is derived from ploidy-normalized copy number data.
24. The method of claim 22 or claim 23, wherein change point copy number feature is assessed across: (i) the telomeric portion of the genome; (ii) the centromeric portion of the genome; or (iii) both the telomeric portion and the centromeric portion of the genome.
25. The method of any one of claims 22-24, wherein the change point copy number feature is a binned feature.
26. The method of any one of claims 1-25, wherein the plurality of features comprise a segment copy number feature, wherein segment copy number is based on the copy number of each genome segment.
27. The method of claim 26, wherein the segment copy number feature is assessed across: (i) the telomeric portion of the genome; (ii) the centromeric portion of the genome; or (iii) both the telomeric portion and the centromeric portion of the genome.
28. The method of claim 26 or claim 27, wherein the segment copy number feature is derived from ploidy-normalized copy number data.
29. The method of any one of claims 26-28, wherein the segment copy number feature is a binned feature.
30. The method of any one of claims 1-29, wherein the plurality of features comprise a breakpoint count per chromosome arm feature in the genome of the tumor of the subject.

31. The method of claim 30, wherein the breakpoint count per chromosome arm feature is assessed across: (i) the telomeric portion of the genome; (ii) the centromeric portion of the genome; or (iii) both the telomeric portion and the centromeric portion of the genome.
32. The method of claim 30 or claim 31, wherein the breakpoint count per chromosome arm feature is a binned feature.
33. The method of any one of claims 1-32, wherein the plurality of features comprise a number of segments with oscillating copy number feature.
34. The method of claim 33, wherein the number of segments with oscillating copy number feature is based on the number of repeated alternating segments between two copy numbers across the genome of the tumor of the subject.
35. The method of claim 33 or claim 34, wherein number of segments with oscillating copy number feature is assessed across: (i) the telomeric portion of the genome; (ii) the centromeric portion of the genome; or (iii) both the telomeric portion and the centromeric portion of the genome.
36. The method of any one of claims 33-35, wherein the number of segments with oscillating copy number feature is a binned feature.
37. The method of any one of claims 1-36, wherein the one or more copy number features comprise a segment minor allele frequency (segMAF) feature, wherein segMAF is based on the minor allele frequency at heterozygous single nucleotide polymorphisms.
38. The method of claim 37, wherein segMAF is assessed across: (i) the telomeric portion of the genome; (ii) the centromeric portion of the genome; or (iii) both the telomeric portion and the centromeric portion of the genome.
39. The method of claim 37 or claim 38, wherein the segment minor allele frequency feature is a binned feature.

40. The method of any one of claims 1-39, wherein the one or more copy number features comprise a number of sequencing reads feature.
41. The method of claim 40, wherein the number of sequencing reads feature is a binned feature.
42. The method of any one of claims 1-41, wherein the plurality of features further comprise a measure of genome-wide loss of heterozygosity of the genome of the tumor of the subject.
43. The method of any one of claims 1-42 wherein the plurality of features comprise one or more short variant features.
44. The method of claim 43, wherein the one or more short variant features comprise at least one of a deletions in microhomology or repetitive regions feature and a mutational signature derived from two or more short variant features.
45. The method of claim 44, wherein the deletions in microhomology or repetitive regions feature are deletions of at least 5 basepairs.
46. The method of any one of claims 1-45, wherein training the HRD model comprises:  
receiving, by the one or more processors, an HRD-positive training dataset, wherein the HRD-positive training dataset comprises a plurality of features associated with an HRD-positive tumor and an HRD-positive label;  
receiving, by the one or more processors, an HRD-negative training dataset, wherein the HRD-negative training dataset comprises a plurality of features associated with an HRD-negative tumor and an HRD-negative label;  
training, by the one or more processors, the HRD model using the HRD-positive training dataset and the HRD-negative training dataset.
47. The method of any one of claims 1-46, further comprising testing, by the one or more processors, the trained model using a HRD-positive testing dataset comprising a HRD-positive control derived from a genome sequence comprising loss-of-function mutations in *BRCA1*, *BRCA2*, both *BRCA1* and *BRCA2*, or biallelic mutations of *BRCA1* and *BRCA2*.

48. The method of any one of claims 1-47, further comprising testing, by the one or more processors, the trained model using a HRD-positive testing dataset comprising a HRD-positive control derived from a genome sequence comprising loss-of-function mutations in at least one of *ATM*, *BARD1*, *BRIP1*, *CDK12*, *CHEK1*, *CHEK2*, *FANCL*, *PALB2*, *RAD51B*, *RAD51C*, *RAD51D*, or *RAD45L*.

49. The method of any one of claims 1-48, further comprising testing, by the one or more processors, the trained model using a HRD-negative testing dataset comprising a HRD-negative training dataset comprising a HRD-negative control derived from a consensus human genome sequence.

50. The method of any one of claims 46-49, wherein training comprises using a HRD-positive training dataset and an HRD-negative training dataset.

51. The method of claim 50, comprising balancing, by the one or more processors, the HRD-positive training dataset and the HRD-negative training dataset prior to training the HRD model.

52. The method of any one of claims 1-51, wherein the tumor in the subject is a prostate cancer, ovarian cancer, breast cancer, non-small cell lung cancer (NSCLC), colorectal cancer (CRC), or pancreatic cancer.

53. The method of any one of claims 1-52, wherein training the HRD model comprises fitting the HRD model to sample data associated with ovarian cancer, non-small cell lung cancer (NSCLC), colorectal cancer (CRC), breast cancer, pancreatic cancer, or prostate cancer, wherein the sample data comprises the subset of the plurality of features.

54. The method of any one of claims 1-53, wherein the tumor is obtained from a sample that is a solid tissue biopsy sample.

55. The method of claim 54, wherein the solid tissue biopsy sample is a formalin-fixed paraffin-embedded (FFPE) sample.

56. The method of any one of claims 1-53, wherein the tumor is obtained from a sample that is a liquid biopsy sample comprising circulating tumor DNA (ctDNA).
57. The method of any one of claims 1-53, wherein the tumor is obtained from a sample that is a liquid biopsy sample comprising cell-free DNA (cfDNA).
58. The method of any one of claims 1-57, further comprising: determining, identifying, or applying the output of the tumor as HRD-positive or HRD-negative as a diagnostic value associated with the patient.
59. The method of any one of claims 1-58, further comprising generating a genomic profile for the subject based on the output of the tumor as HRD-positive or HRD-negative.
60. The method of claim 59, further comprising administering an anti-cancer agent or applying an anti-cancer treatment to the subject based on the generated genomic profile.
61. The method of any one of claims 1-60, wherein the output of the tumor as HRD-positive or HRD-negative is used in generating a genomic profile for the subject.
62. The method of any one of claims 1-61, wherein the output of the tumor as HRD-positive or HRD-negative is used in making suggested treatment decisions for the subject.
63. The method of any one of claims 1-62, wherein the output of the tumor as HRD-positive or HRD-negative is used in applying or administering a treatment to the subject.
64. The method of any one of claims 1-63, wherein the HRD model is a machine learning model.
65. The method of any one of claims 1-64, wherein the subject has a cancer, is at risk of having a cancer, or is suspected of having a cancer.
66. A method of treating cancer in a subject, comprising:  
(a) identifying the tumor as HRD-positive or HRD-negative according to the method of any one of claims 1-65;

(b) administering to the subject a therapeutically effective amount of a drug effective in a HRD positive tumor if the tumor of the cancer is assessed as HRD positive.

67. The method of claim 66, wherein the drug effective in a HRD positive tumor is a platinum-based drug or a PARP inhibitor.

68. The method of claim 66 comprising administering to the subject a therapeutically effective amount of a drug that is not a platinum-based drug or a PARP inhibitor if the tumor is assessed as HRD negative.

69. A method for selecting a therapy for a cancer in a subject, the method comprising:  
(a) assessing a tumor of the cancer as HRD-positive or HRD-negative according to the method of any one of claims 1-65;  
(b) selecting a therapy that is effective in a HRD positive tumor if the cancer is assessed as HRD positive.

70. The method of claim 69, comprising selecting a therapy that is not a platinum-based drug or a PARP inhibitor if the tumor is assessed as HRD negative.

71. The method of claim 70, wherein the therapy that is effective in a HRD positive tumor is a platinum-based drug or a PARP inhibitor.

72. A computer system, comprising:  
one or more processors;  
a memory; and  
one or more programs, wherein the one or more programs are stored in the memory and configured to be executed by the one or more processors, the one or more programs including instructions for: performing the method of any one of claims 1-65.

73. A non-transitory computer-readable storage medium storing one or more programs, the one or more programs comprising instructions, which when executed by one or more processors of an electronic device, cause the electronic device to perform the method of any one of claims 1-65.

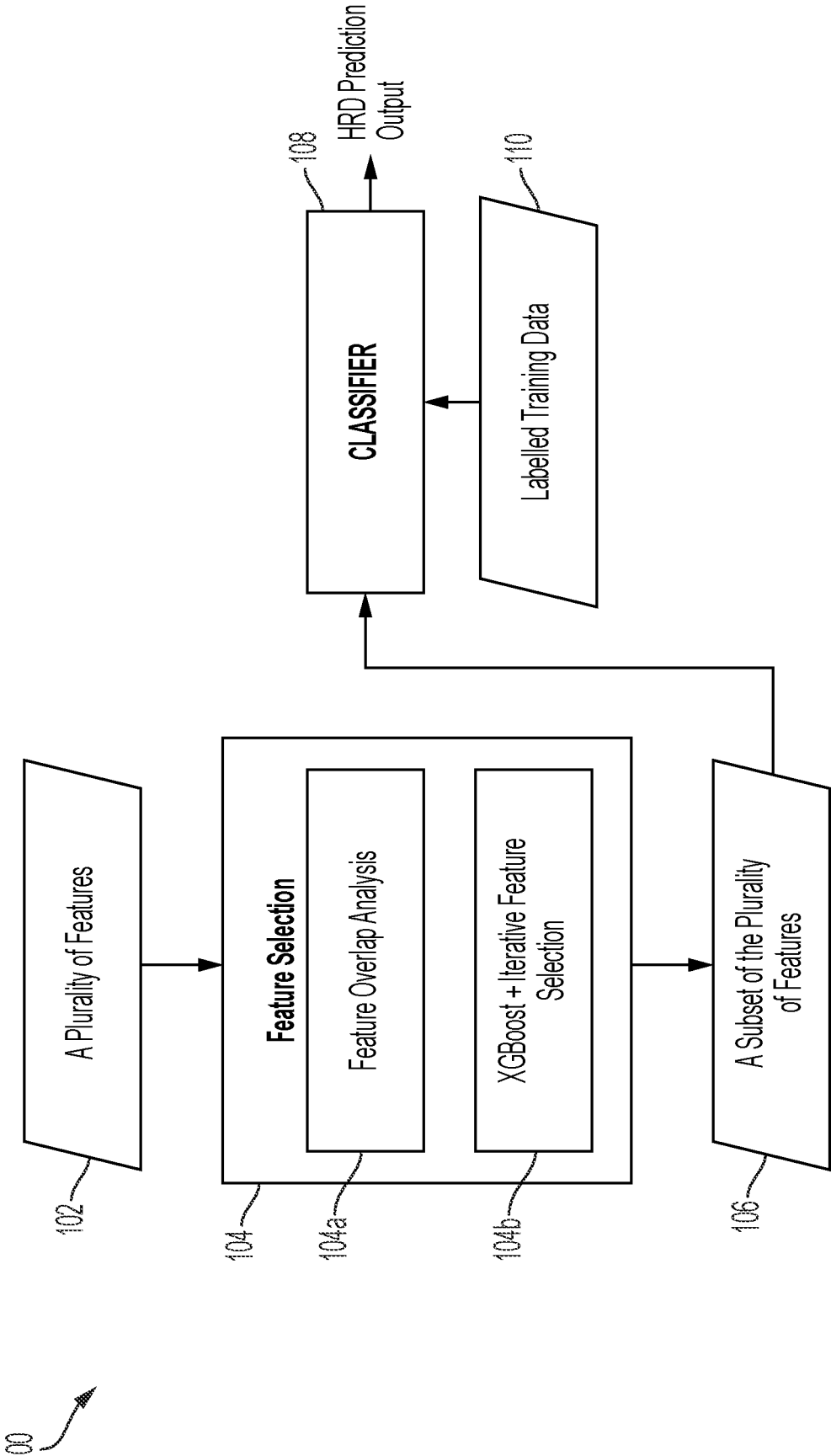


FIG. 1

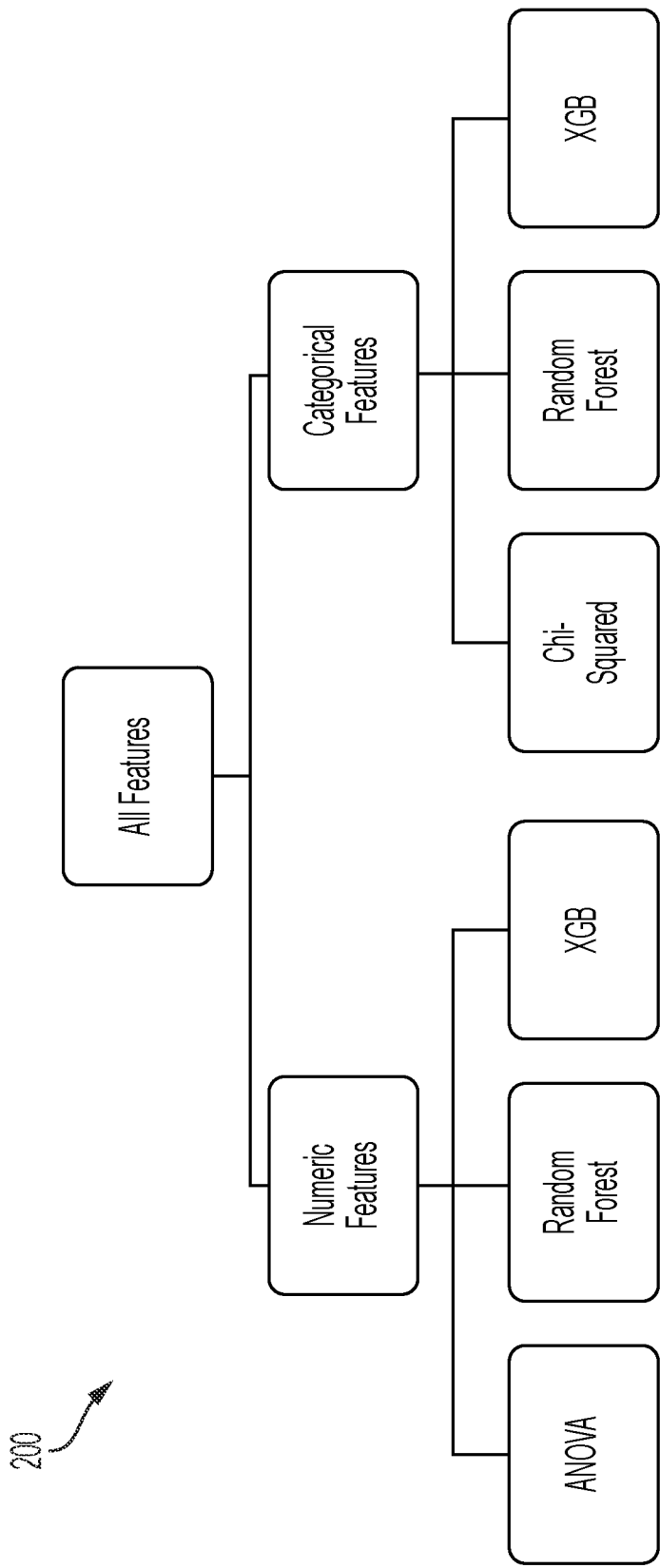


FIG. 2

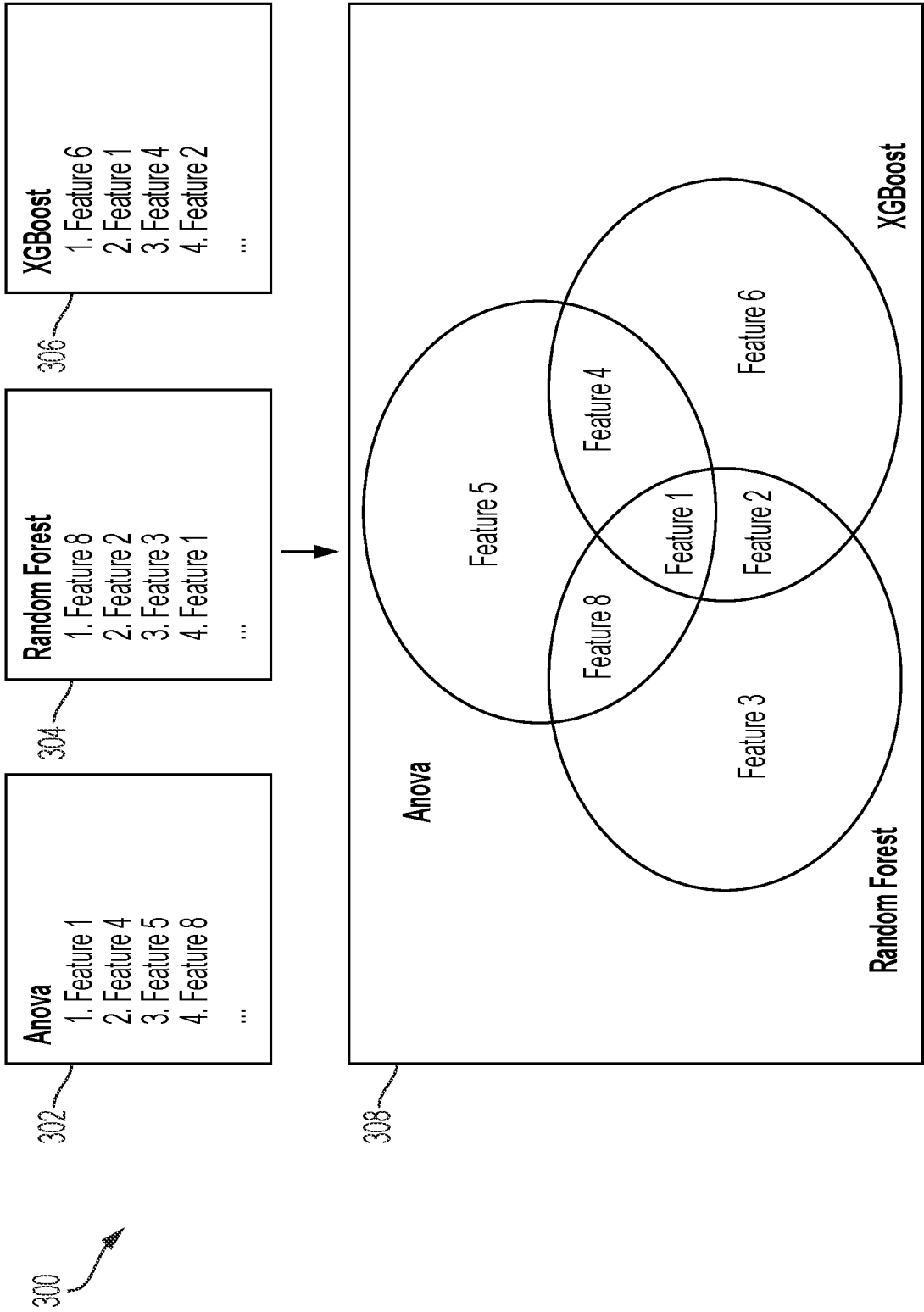


FIG. 3A

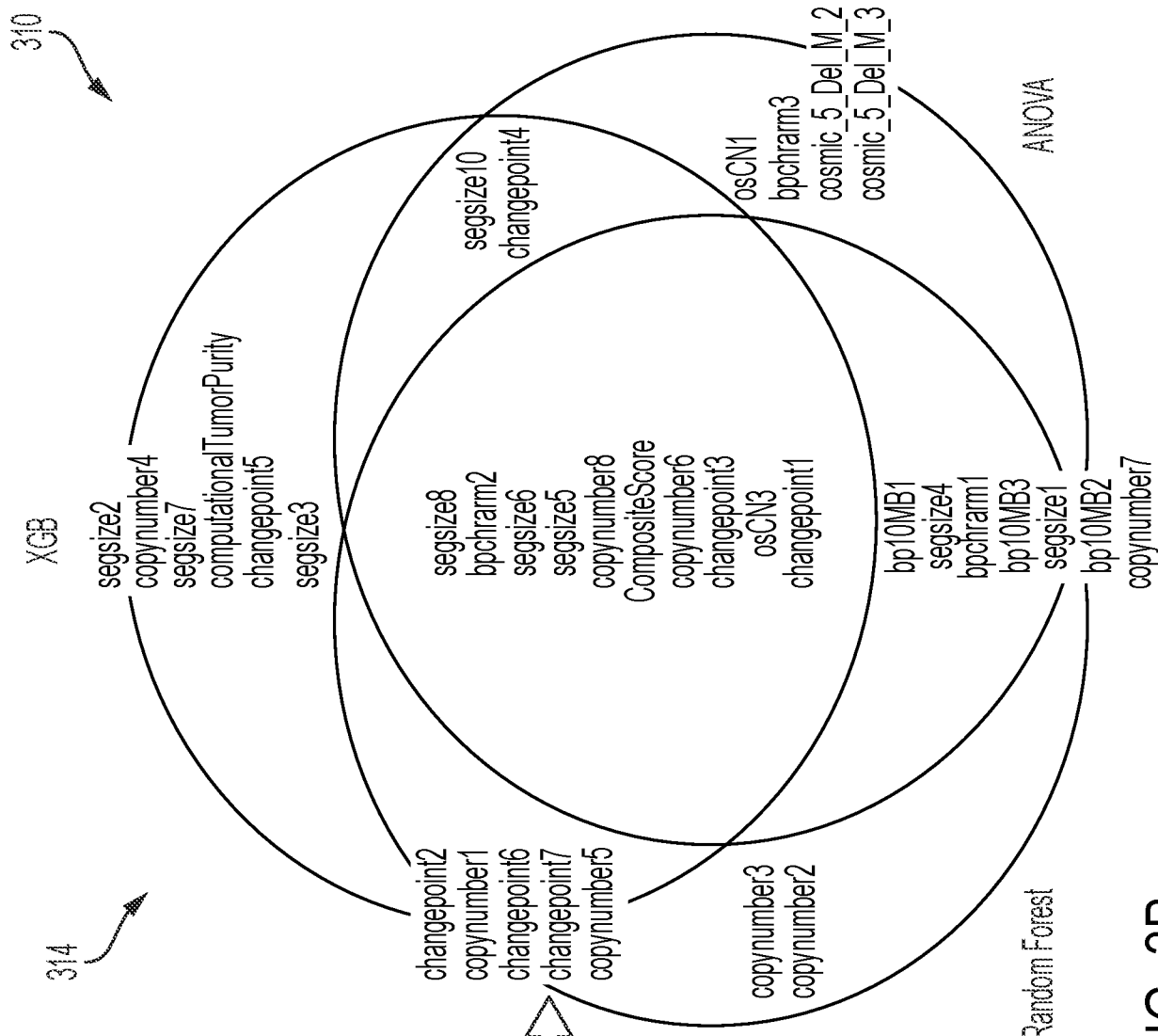
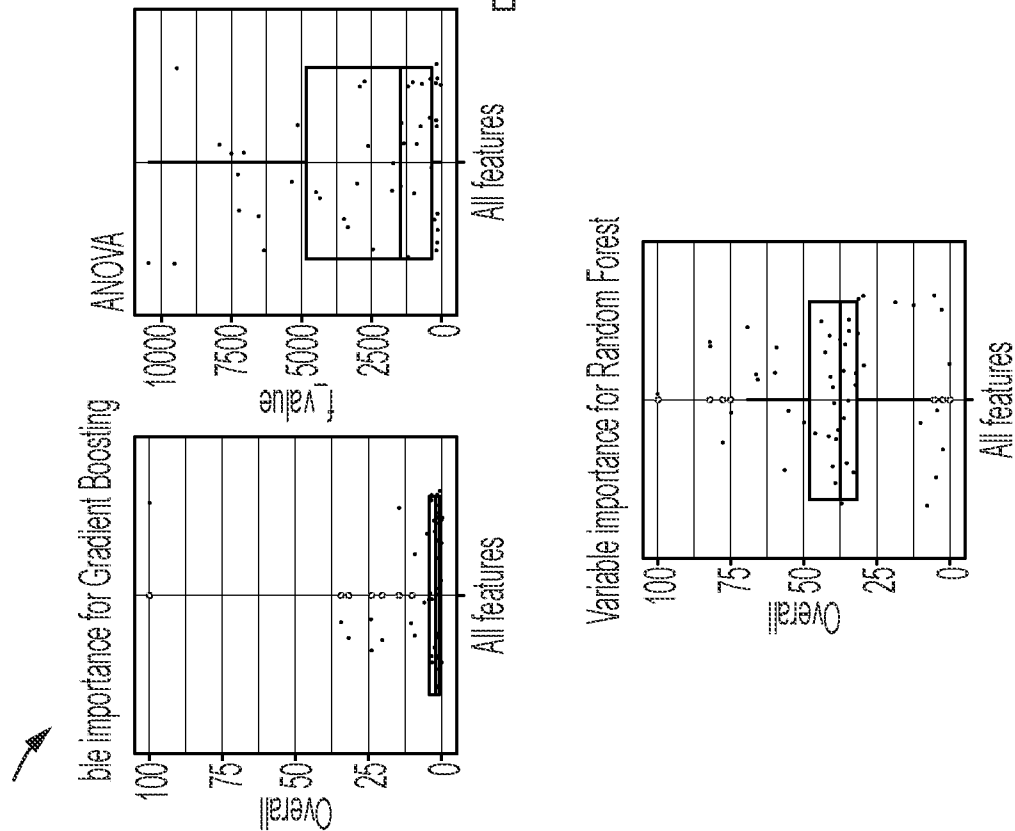


FIG. 3B



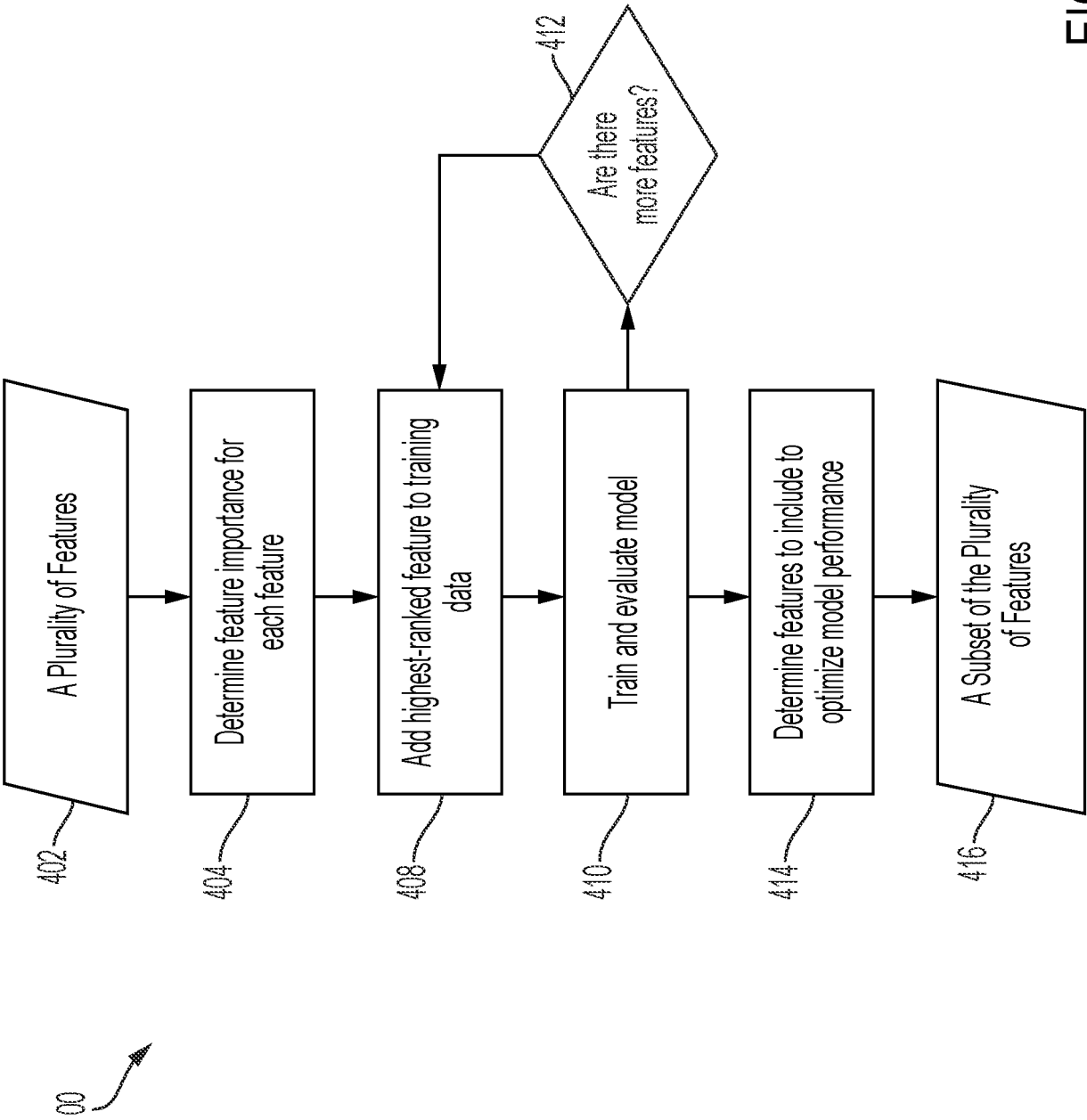
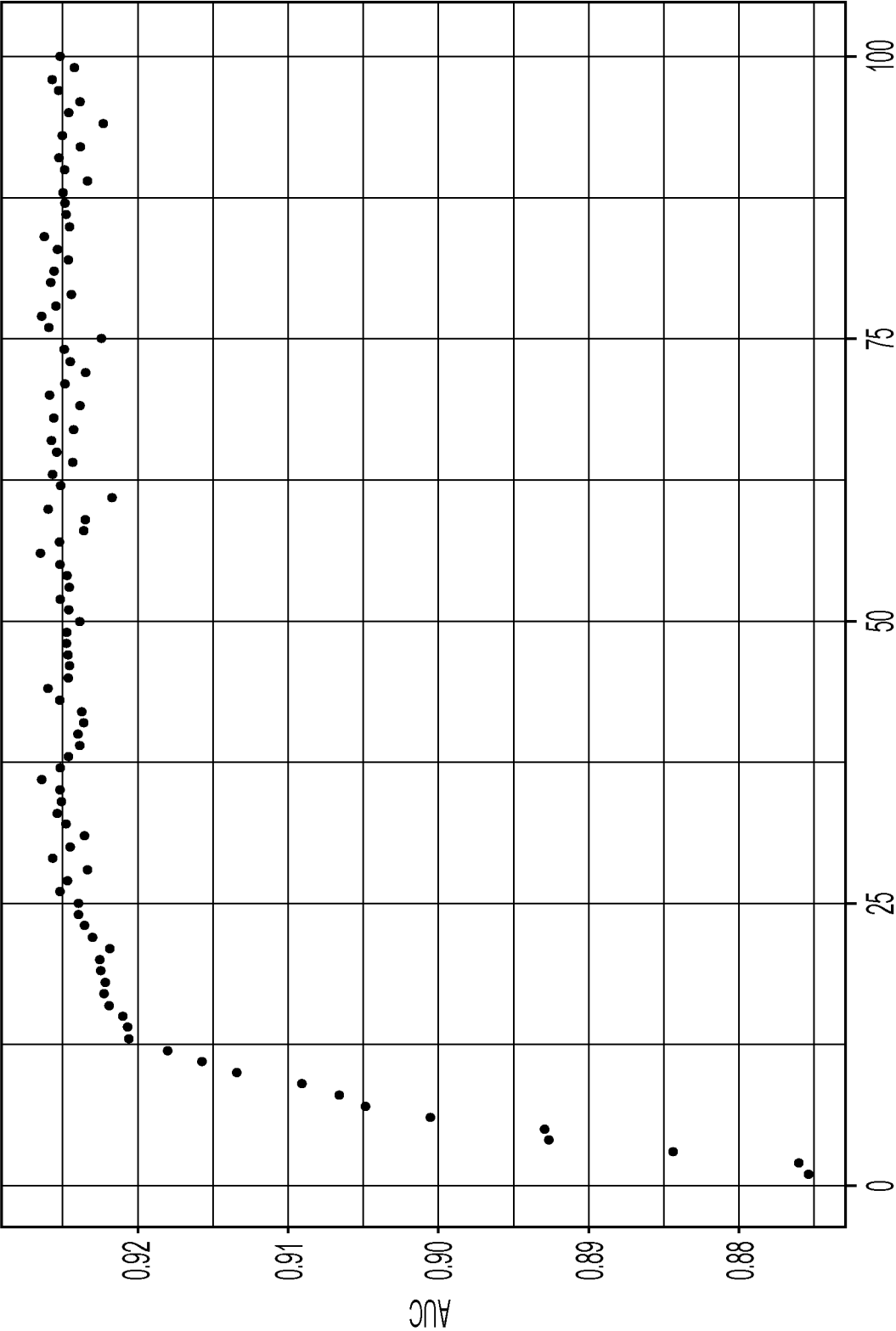


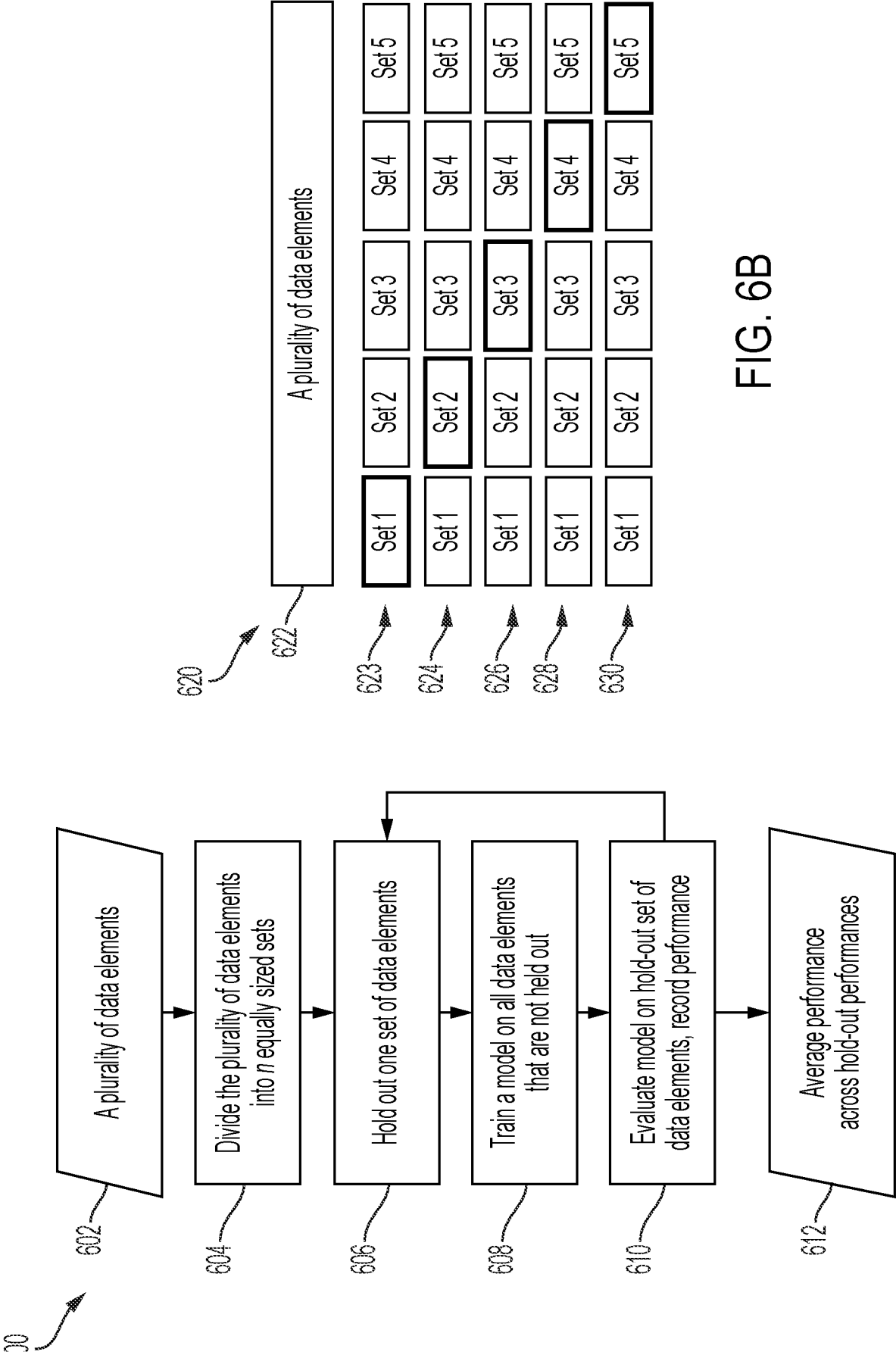
FIG. 4

500



Number of Features

FIG. 5



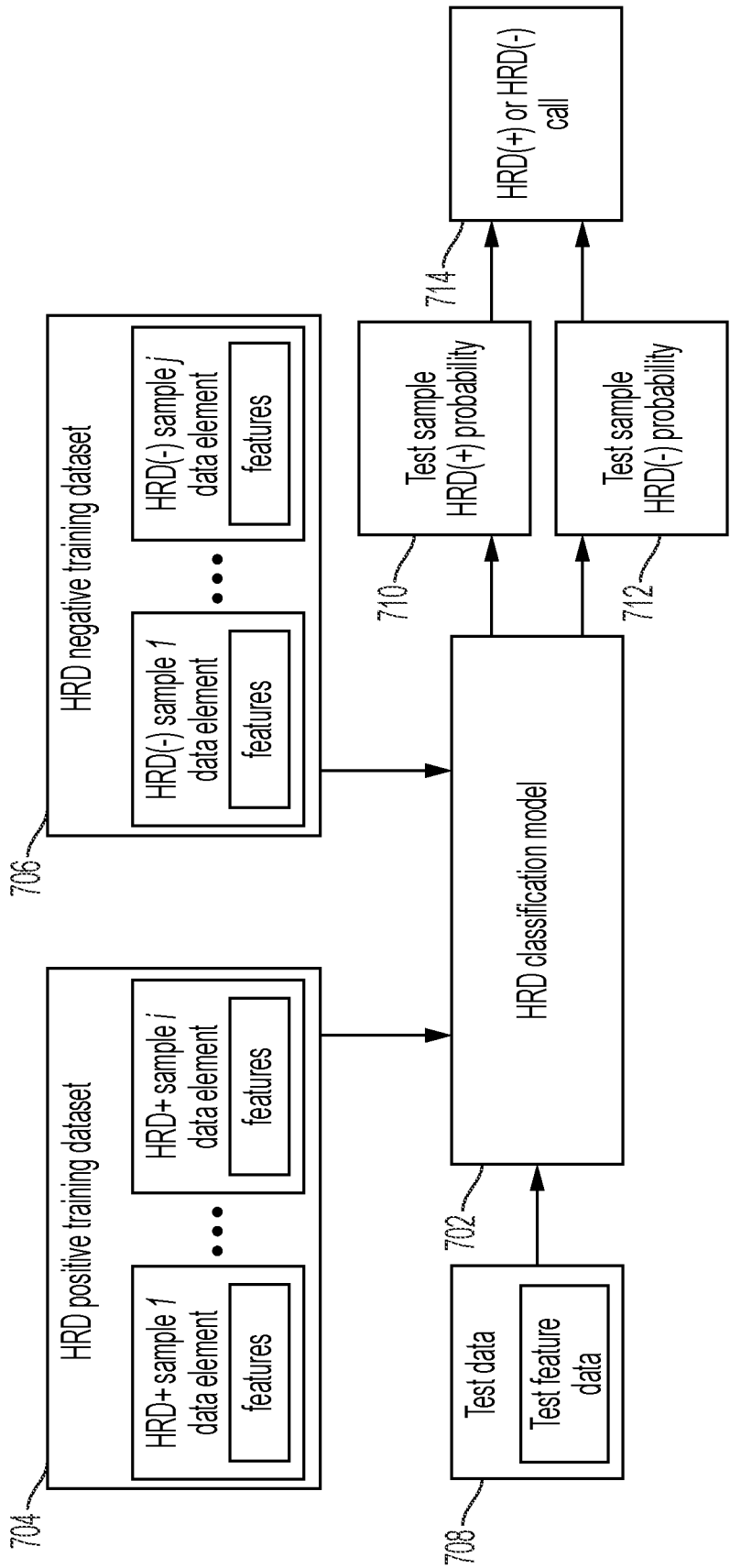


FIG. 7

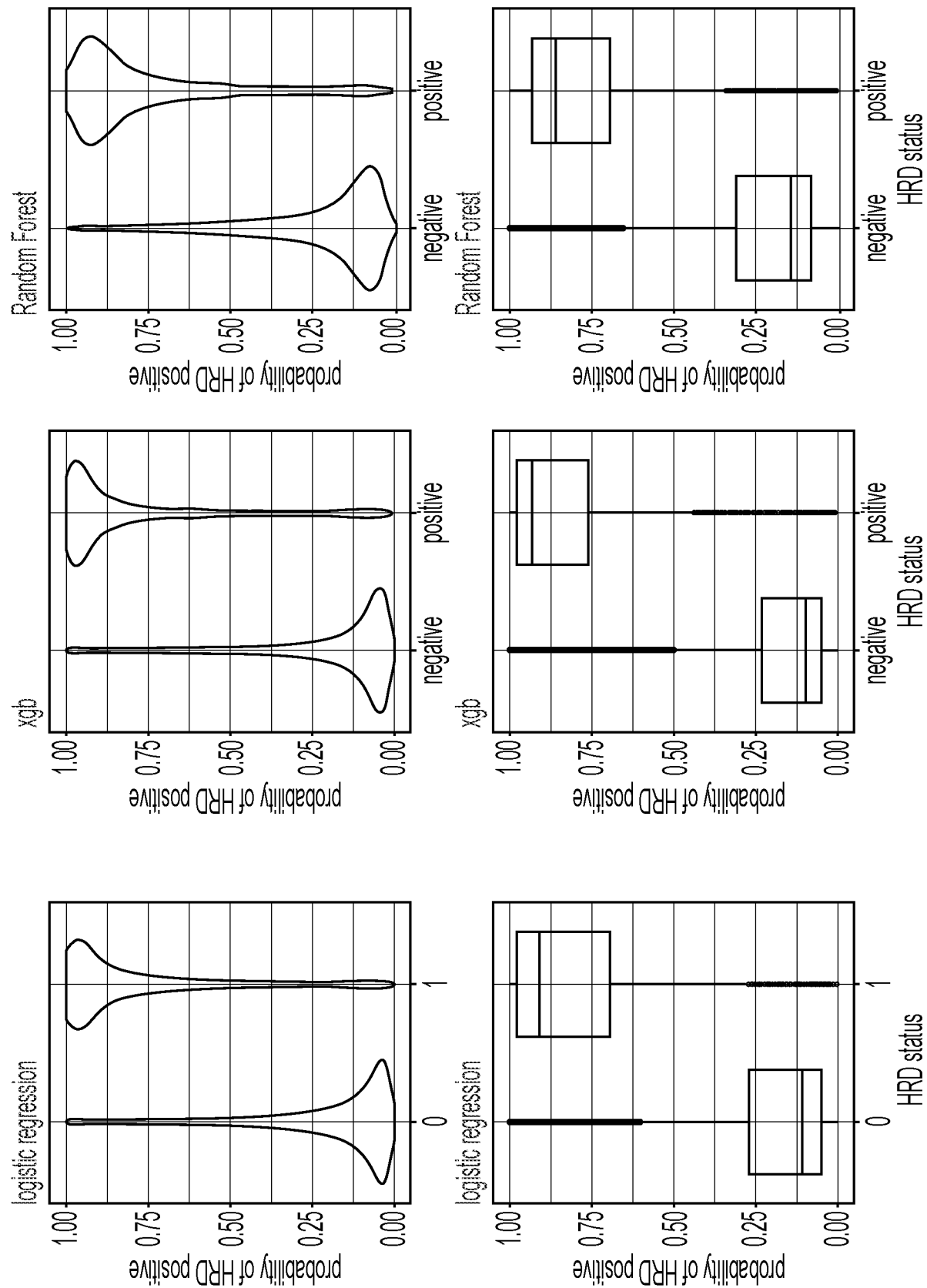
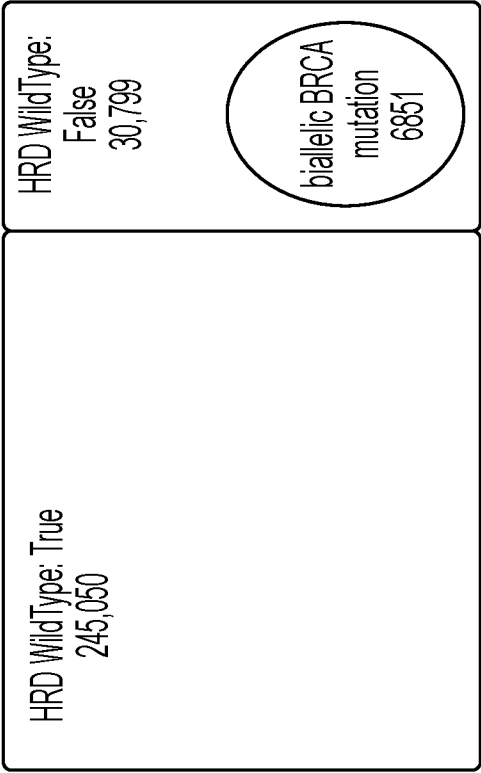
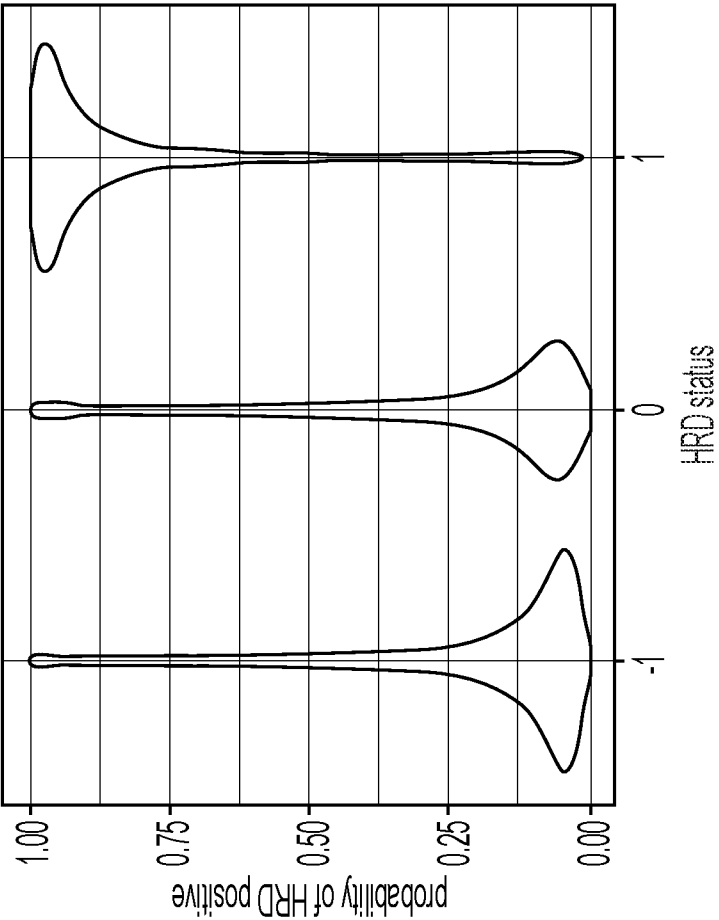


FIG. 8



- HRD WildType True: Specimens **without** mutations in a basket of 14 HRD genes\*
- HRD WildType False: Specimens **with** mutations in at least one HRD genes\*
- Biallelic BRCA mutation: Specimens with biallelic BRCA1/2 mutation (homozygous SV, multiple SV, homozygous deletion)

FIG. 9

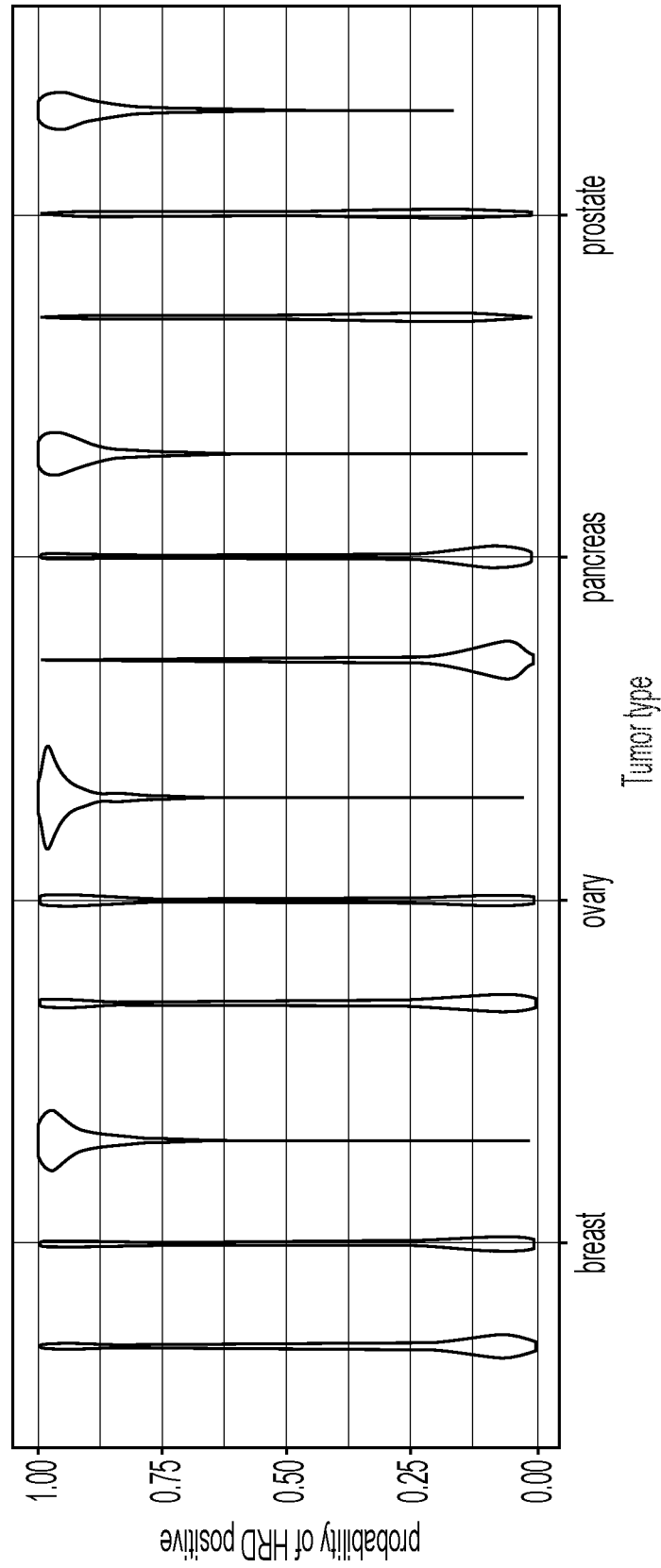


FIG. 10

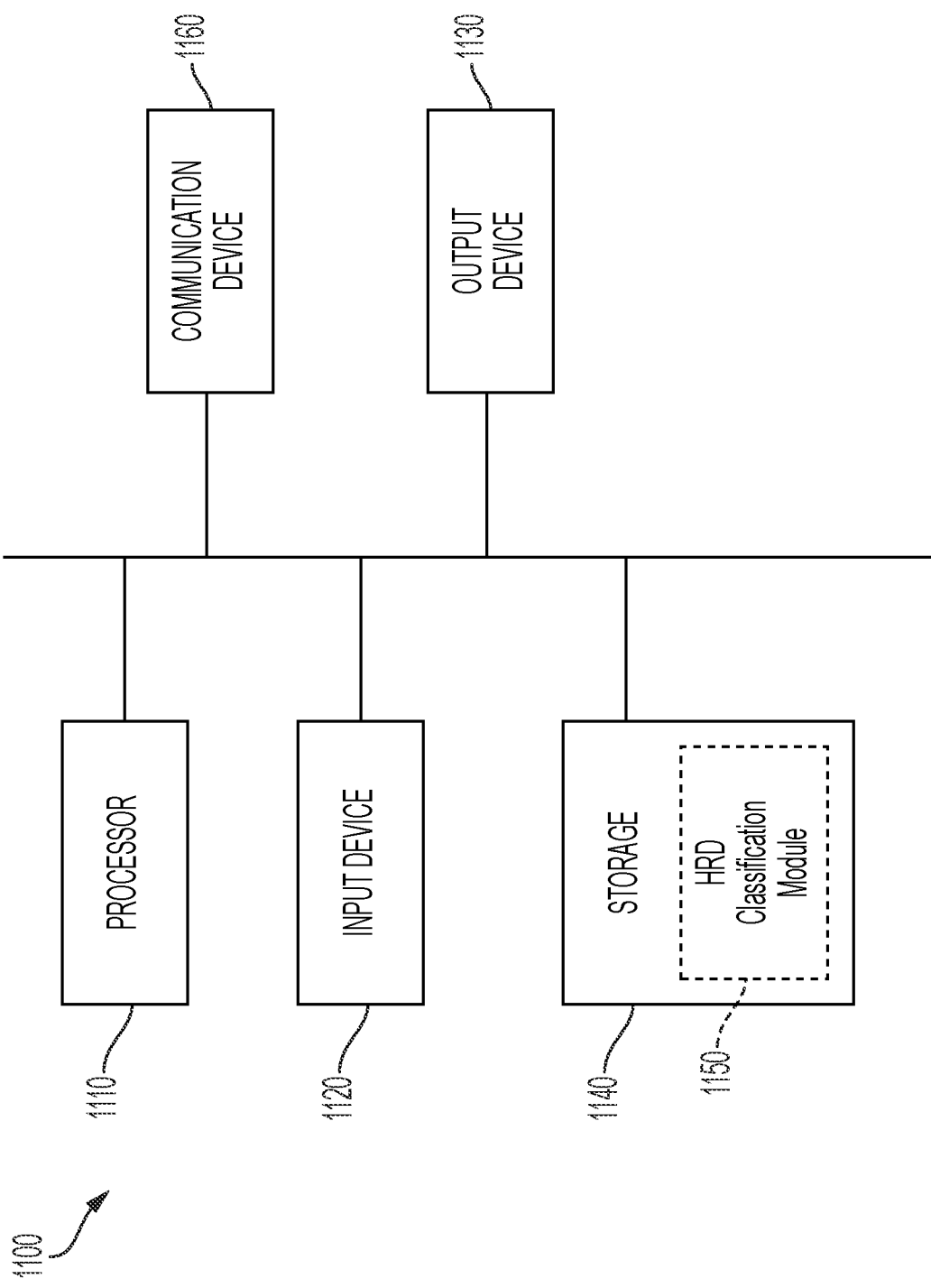


FIG. 11

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/US22/73167

## A. CLASSIFICATION OF SUBJECT MATTER

IPC - INV. G16B 20/10; C12Q 1/6886; G06F 17/18; G16B 20/20; G16B 40/20 (2022.01)

ADD.

CPC - INV. G16B 20/10; C12Q 1/6886; G06F 17/18; G16B 20/20; G16B 40/20

ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

See Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History document

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                                                          | Relevant to claim No. |
|-----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| X         | US 2020/0255909 A1 (TEMPUS) 13 August 2020; figures 1A-1B and 2; paragraphs [0047, 0051, 0057, 0061, 0064, 0068, 0069, 0072, 0081-0082, 0092, 0113-0115, 0127, 0147, 0149]. | 1-4                   |
| A         | US 2021/0172024 A1 (TEMPUS) 10 June 2021; entire document.                                                                                                                  | 1-4                   |
| A         | US 2018/0030544 A1 (MYRIAD GENETICS INC) 01 February 2018; entire document                                                                                                  | 1-4                   |
| A         | US 2021/0166785 A1 (TEMPUS LABS, INC .,) 03 June 2021; entire document                                                                                                      | 1-4                   |

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

\* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"D" document cited by the applicant in the international application

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&amp;" document member of the same patent family

Date of the actual completion of the international search

18 August 2022 (18.08.2022)

Date of mailing of the international search report

SEP 06 2022

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents  
P.O. Box 1450, Alexandria, Virginia 22313-1450  
Facsimile No. 571-273-8300

Authorized officer

Shane Thomas

Telephone No. PCT Helpdesk: 571-272-4300

# INTERNATIONAL SEARCH REPORT

International application No.

PCT/US22/73167

## Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:  
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☐ Claims Nos.:  
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☒ Claims Nos.: 5-73  
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

## Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

### Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.