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

SUMMARY

[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. In one aspect, the present disclosure refers to 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.

[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. In one aspect, the present disclosure refers to 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.

[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. In one aspect, the present disclosure refers to 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.

[0008] In some examples 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 examples 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 examples 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 examples 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 examples 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 examples, the classification result comprises at least one of a HRD-positive likelihood score and a HRD-negative likelihood score. In some examples, 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 examples, 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 examples of the described methods, the HRD model is a classification model, a regression model, a neural network, or any combination thereof. In some examples, 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 examples, 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 examples 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 examples of the described methods, at least one of the plurality of features is assessed across the centromeric portion of the genome. In some examples of the described methods, at least one of the plurality of features is assessed across the telomeric portion of the genome.

[0014] In some examples 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 examples 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 examples, 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 examples, *x* is between about 1 and about 100 megabases. In some examples, *x* is about 10 megabases, about 25 megabases, about 50 megabases, or about 100 megabases. In some examples, the breakpoint count per *x* megabases feature is a binned feature.

[0016] In some examples 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 examples, the change point copy number feature is derived from ploidy-normalized copy number data. In some examples, 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 examples, the change point copy number feature is a binned feature.

[0017] In some examples 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 examples, 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 examples, the segment copy number feature is derived from ploidy-normalized copy number data. In some examples, the segment copy number feature is a binned feature.

[0018] In some examples 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 examples, 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 examples, the breakpoint count per chromosome arm feature is a binned feature.

[0019] In some examples of the described methods, the plurality of features comprise a number of segments with oscillating copy number feature. In some examples, 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 examples, 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 examples, the number of segments with oscillating copy number feature is a binned feature.

[0020] In some examples 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 examples, 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 examples, the segMAF feature is a binned feature.

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

[0022] In some examples 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 examples of the described methods, the plurality of features comprise one or more short variant features. In some examples, 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 examples, the deletions in microhomology or repetitive regions feature are deletions of at least 5 basepairs.

[0024] In some examples 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 examples, training comprises using a HRD-positive training dataset and an HRD-negative training dataset. In some examples, 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 examples 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 examples, training comprises using a HRD-positive training dataset and an HRD-negative training dataset. In some examples, 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 examples 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 examples, training

comprises using a HRD-positive training dataset and an HRD-negative training dataset. In some examples, 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 examples 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 examples, training comprises using a HRD-positive training dataset and an HRD-negative training dataset. In some examples, 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 examples 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 examples 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 examples of the described methods, the tumor is obtained from a sample that is a solid tissue biopsy sample. In some examples, the solid tissue biopsy sample is a formalin-fixed paraffin-embedded (FFPE) sample. In some examples of the described methods, the tumor is obtained from a sample that is a liquid biopsy sample comprising circulating tumor DNA (ctDNA). In some examples 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 examples 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 examples 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 examples, 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 examples 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 examples 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 examples 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 examples of the described methods, the HRD model is a machine learning model.

[0033] In some examples 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 one aspect, the present disclosure refers to use of a therapeutically effective amount of a drug effective in a HRD-positive tumor in the manufacture of a medicament for treating a tumor in a subject, wherein: (a) the tumor is to be identified as HRD-positive or HRD-negative according to the method as disclosed herein; and (b) if the tumor is assessed as HRD positive, a therapeutically effective amount of a drug effective in a HRD positive tumor is to be administered. In some examples, the drug effective in a HRD positive tumor is a platinum-based drug or a PARP inhibitor. In some examples, 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 one aspect, the present disclosure refers to a method for selecting a therapy for a tumor to be administered to a subject, comprising: (a) assessing the tumor as HRD-positive or HRD-negative according to the method as disclosed herein; and (b) selecting a therapy that is effective in a HRD-positive tumor if the tumor is assessed as HRD positive. In some examples, 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 examples, 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.

CLAIMS

In the claims

1. 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.

2. 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.

3. 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.

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 claim 2, wherein the classifying comprises determining at least one of a HRD-positive likelihood score and a HRD-negative likelihood score.

9. The method of claim 8, comprising recording, in a digital electronic file associated with the subject, at least one of the HRD-positive likelihood score and the HRD-negative likelihood score.

10. The method of any one of claims 8-9, comprising recording in a digital electronic file associated with the subject, a designation 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.

11. The method of any one of claims 1-10, 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.

12. The method of any one of claims 1-11, 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.

13. The method of any one of claims 1-12, 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.

14. The method of any one of claims 1-13, wherein the output of the tumor as HRD-positive or HRD-negative is used in making suggested treatment decisions for the subject.

15. The method of any one of claims 1-14, wherein the HRD model is a machine learning model.

16. Use of a therapeutically effective amount of a drug effective in a HRD-positive tumor in the manufacture of a medicament for treating a tumor in a subject, wherein:

(a) the tumor is to be identified as HRD-positive or HRD-negative according to the method of any one of claims 1-15; and

(b) if the tumor is assessed as HRD positive, a therapeutically effective amount of a drug effective in a HRD positive tumor is to be administered.

17. The use of claim 16, wherein the drug effective in a HRD positive tumor is a platinum-based drug or a PARP inhibitor.

18. The use of claim 16, wherein a therapeutically effective amount of a drug that is not a platinum-based drug or a PARP inhibitor is to be administered if the tumor is assessed as HRD negative.

19. A method for selecting a therapy for a tumor to be administered to a subject, comprising:

(a) assessing the tumor as HRD-positive or HRD-negative according to the method of any one of claims 1-15; and

(b) selecting a therapy that is effective in a HRD-positive tumor if the tumor is assessed as HRD positive.

20. The method of claim 19, comprising selecting a therapy that is not a platinum-based drug or a PARP inhibitor if the tumor is assessed as HRD negative.

21. The method of claim 20, wherein the therapy that is effective in a HRD positive tumor is a platinum-based drug or a PARP inhibitor.