





Copy number alteration features in pan-cancer homologous recombination deficiency prediction and biology

Huizi Yao ^{1,2,3,5}, Huimin Li^{1,2,3,5}, Jinyu Wang^{1,2,3,5}, Tao Wu ¹, Wei Ning¹, Kaixuan Diao¹, Chenxu Wu¹, Guangshuai Wang¹, Ziyu Tao ¹, Xiangyu Zhao¹, Jing Chen¹, Xiaoqin Sun¹ & Xue-Song Liu ^{1,4}✉

Homologous recombination deficiency (HRD) renders cancer cells vulnerable to unrepaired double-strand breaks and is an important therapeutic target as exemplified by the clinical efficacy of poly ADP-ribose polymerase (PARP) inhibitors as well as the platinum chemotherapy drugs applied to HRD patients. However, it remains a challenge to predict HRD status precisely and economically. Copy number alteration (CNA), as a pervasive trait of human cancers, can be extracted from a variety of data sources, including whole genome sequencing (WGS), SNP array, and panel sequencing, and thus can be easily applied clinically. Here we systematically evaluate the predictive performance of various CNA features and signatures in HRD prediction and build a gradient boosting machine model (HRD_{CNA}) for pan-cancer HRD prediction based on these CNA features. CNA features BP10MB[1] (The number of breakpoints per 10MB of DNA is 1) and SS[> 7 & ≤ 8] (The log10-based size of segments is greater than 7 and less than or equal to 8) are identified as the most important features in HRD prediction. HRD_{CNA} suggests the biallelic inactivation of *BRCA1*, *BRCA2*, *PALB2*, *RAD51C*, *RAD51D*, and *BARD1* as the major genetic basis for human HRD, and may also be applied to effectively validate the pathogenicity of *BRCA1/2* variants of uncertain significance (VUS). Together, this study provides a robust tool for cost-effective HRD prediction and also demonstrates the applicability of CNA features and signatures in cancer precision medicine.

¹School of Life Science and Technology, ShanghaiTech University, Shanghai, China. ²Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai, China. ³University of Chinese Academy of Sciences, Beijing, China. ⁴Shanghai Clinical Research and Trial Center, Shanghai, China. ⁵These authors contributed equally: Huizi Yao, Huimin Li, Jinyu Wang. ✉email: liuxs@shanghaitech.edu.cn

DNA damage is a common hallmark of cancer, and DNA double-strand break (DSB) is the most hazardous type of DNA damage which can be repaired by error-free homologous recombination (HR) pathway¹. In cancer, accurate detection of homologous recombination deficiency (HRD) is of clinical relevance as HRD tumors are reported to be sensitive to poly ADP-ribose polymerase (PARP) inhibitors², as well as to platinum chemotherapy drugs³.

Germline BRCA genes testing⁴ and the genomic instability score (GIS) are the currently widely used HR status diagnostic schemes. Germline BRCA testing could overlook the epigenetic silencing^{5–8}, and other pathogenic somatic mutations⁹, and at the same time strongly relies on the completeness and accuracy of clinical variant annotation databases. GIS, also named as “HRD score”, mainly consists of three SNP array-based methods: telomeric allelic imbalance (TAI)¹⁰, loss of heterozygosity (LOH)¹¹, and large-scale transition (LST)¹². One disadvantage of HRD score is that the optimal threshold is not consistent in patients with different types of cancer^{13–15}. Recently, HRDetect^{16,17} and CHORD⁹ have been constructed for HRD prediction, however, these tools need whole genome sequencing (WGS) or whole exome sequencing (WES) data, which is expensive and lacks clinical applicability.

Copy number alteration (CNA) is an important type of cancer-driving genetic alteration, and CNA signatures and features are emerging cancer-inherent patterns. Recently, patterns and mutational processes of CNA signature have started to be revealed^{18–21}. CNA information can be obtained from a diverse type of data, such as shallow WGS, WES, SNP array, and panel sequencing, and could represent a cost-effective type of biomarker for cancer diagnosis and clinical response prediction. However, the efficacy and application of CNA signatures and features in HRD prediction have not been established yet.

Here we develop a robust HRD predictor HRD_{CNA} (Homologous recombination deficiency prediction by copy number alteration features) based on CNA features. HRD_{CNA} model is trained on the CNA data of 1,470 cancer samples generated from WGS or SNP array and is validated with 831 samples generated from WGS, SNP array, shallow WGS, or panel sequencing data. HRD_{CNA} can precisely predict HR status using CNA features data derived from different platforms or different cancer types with consistent cut-off. HRD_{CNA} provides a new direction for effectively validating the pathogenicity of uncertain important variants (VUS) of BRCA1 and BRCA2, and suggests the major genetic basis for human HRD. The applicability of CNA features and signatures in cancer precision medicine is also well demonstrated in HRD_{CNA}.

Results

Data collection for HRD prediction model training and validation. For the development of the HRD prediction model, 1854 samples (WGS data) from the pan-cancer analysis of whole genomes⁹ (PCAWG) and 560 breast cancer samples (SNP array data)¹⁶ are collected. To obtain a high-confidence training dataset of HRD, samples with BRCA1/2 deficiency are screened for classifier training. In total, 130 cancer samples with loss-of-function mutations in BRCA1/2 are labeled as HRD, while 1340 samples without inactivation mutations in known HR genes are labeled as homologous recombination proficiency (HRP). Then these samples are randomly split into training and held-out datasets by the ratio of 8:2 for model training and testing respectively. The performance of the HRD prediction model is further evaluated on three independent validation datasets: a total of 633 cancer samples, including 66 breast cancer samples with WGS, 66 breast cancer samples with SNP array sequencing, and 501 pan-cancer samples with panel sequencing (Fig. 1).

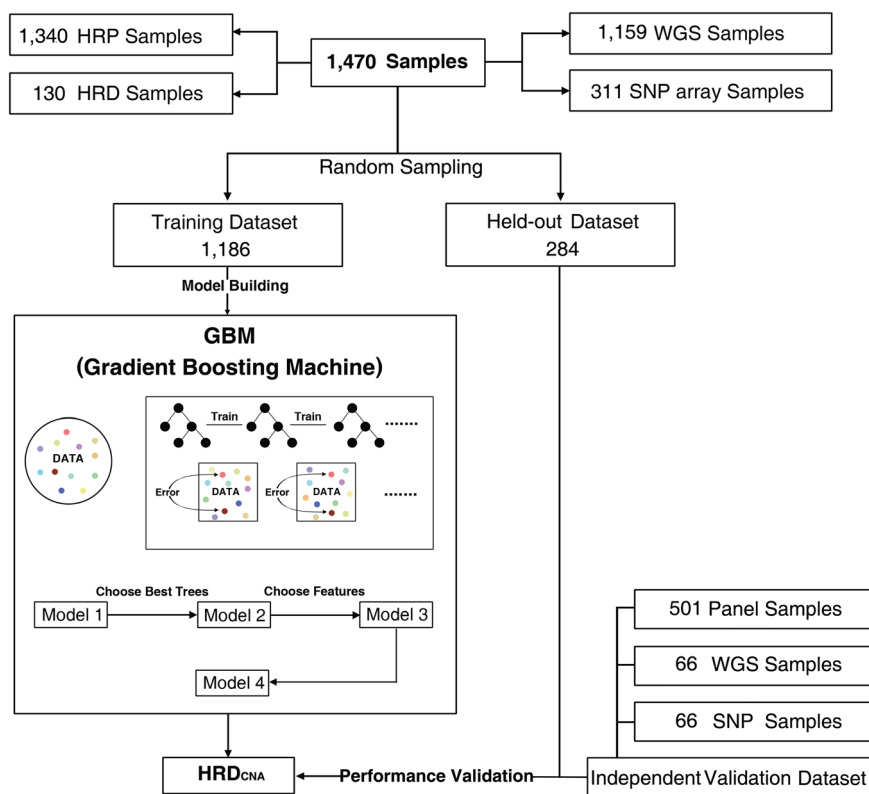


Fig. 1 HRD_{CNA} is a gradient boosting machine for pan-cancer homologous recombination deficiency (HRD) prediction. Training and validation processes of HRD_{CNA}, which outputs the probability of being HRD.

Selection of the method for HRD prediction model building.

To find the best HRD prediction model, a total of 9 machine learning models (see Methods) are trained using the training dataset, and the performances of each machine learning model are reported as the area under the receiver operating characteristic (ROC) curve (AUC) and the area under the precision-recall curves (PR-AUC) in all datasets (Supplementary Fig. 1a–f). Gradient boosting machine (GBM) model shows the best performance in AUC and other metrics like accuracy, precision, and F1 score on the held-out dataset (Supplementary Fig. 1g, h), thus it is selected for subsequent HRD prediction model development (Fig. 1).

HRD_{CNA}, pan-cancer HRD predictor based on CNA features.

CNA signatures and features are emerging cancer inherent indicators. Wang et al.²⁰ classified CNA segments based on 8 types of CNA features, then performed the first CNA signature analysis in prostate cancer using the non-negative matrix factorization (NMF) algorithm with these CNA features, and the CNA signatures developed with this method are named “Sig-CNS” here²⁰. Ruben et al. conducted a pan-cancer CNA signature study and reported 17 signatures (named “Sig-CX” here)¹⁹. CNA features are countable parameters developed in Wang et al. study²⁰. We compare the performance of three different models using CNA features or using the reported two sets of CNA signatures (Sig-CNS and Sig-CX). AUC and PR-AUC results show that HRD prediction models construct with CNA features exhibit improved performance compared with models built with CNA signatures (Fig. 2a, b), thus CNA features are chosen for subsequent HRD model development.

To ensure that the parameters identified are robust and generalizable, 10-fold cross-validation (CV) strategy is used. Then

the relative variable importance is calculated²², and the features with the top 10 relative influence score among all features in the model are selected as the optimal modeling features (Fig. 2c). Then we select the cancer types with higher HRD incidence to analyze the differences in the 10 features (Supplementary Fig. 2). The analysis results show that these features have significant differences in HRD and HRP samples, not only in pan-cancer but also in individual cancers. Finally, we trained an HRD model using these selected CNA features (Fig. 2d). The BP10MB[1] and SS[>7 & <=8] are found to be the top two important predicting factors of HRD (Fig. 2c). The final model is named “HRD_{CNA}”.

HRD_{CNA} model validation and application.

HRD_{CNA} shows excellent performance in the held-out dataset and also three independent validation datasets (Fig. 3a–f). The performance of HRD_{CNA} model is not influenced by the CNA assay platforms (Supplementary Fig. 3). ShallowHRD, a tool to evaluate tumor HRD based on a CNA profile derived from WGS at low sequencing depth²³, shows 0.88 sensitivity and 0.91 specificity for HRD detection. Compared with ShallowHRD, HRD_{CNA} shows consistently improved performance using CNA profiles derived from different sequencing depths of WGS using downsampling (Supplementary Fig. 4). Since HRD cases disproportionately represented in some cancers, it remains to be verified whether HRD_{CNA} model applies to all cancer types. We verified the performance of HRD_{CNA} model in individual cancers separately. The results show that the model had excellent applicability in breast, ovarian, pancreatic lymphatic, prostate, and liver cancers (Supplementary Fig. 5).

We compare the performance of HRD_{CNA} with HRDetect, LOH, TAI, LST, and HRD score. HRD_{CNA} shows improved performance compared with HRDetect in HRD prediction in an

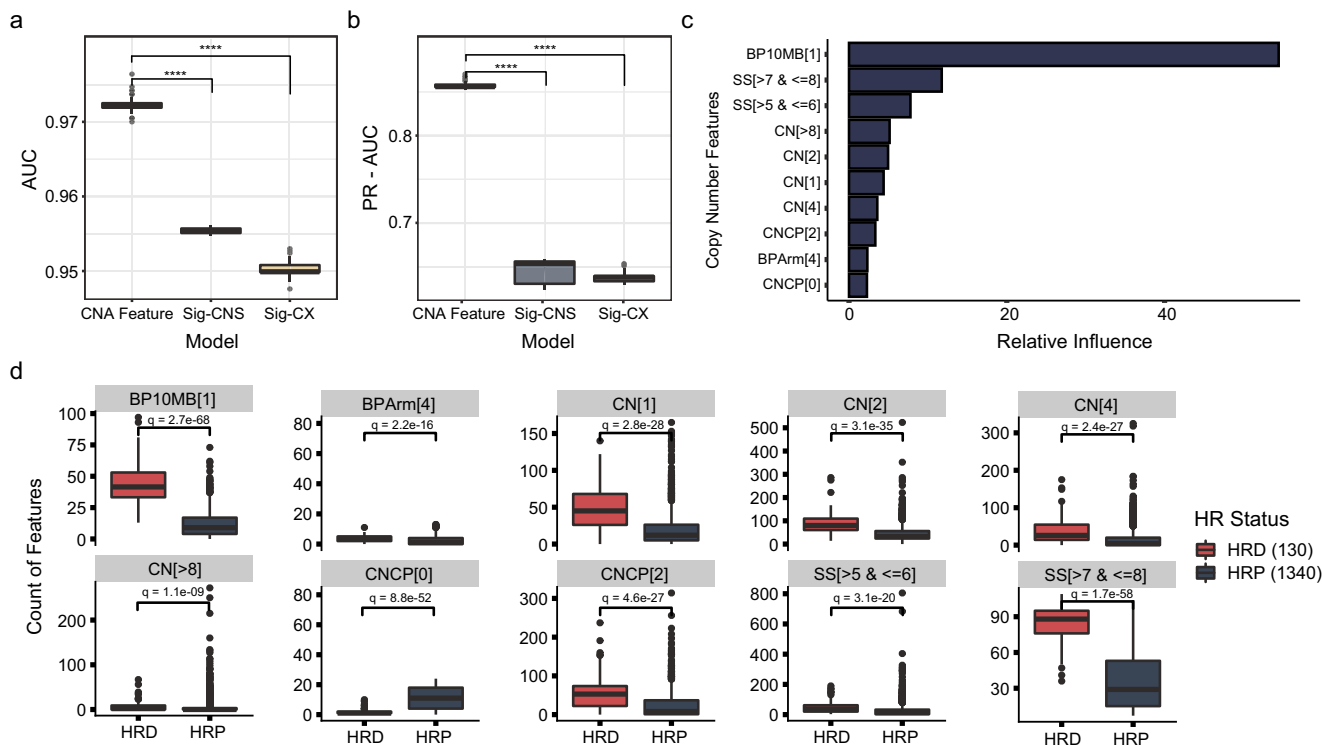


Fig. 2 HRD_{CNA}, choosing CNA features rather than CNA signatures to predict HRD. **a, b** Performance comparisons between the HRD prediction models constructed using CNA features, or the activities of two sets of CNA signatures, Sig-CNS (Method from Wang et al.) or Sig-CX (Method from Ruben et al.). The differences in the area under the receiver operating characteristic curves (AUC) **a** and the area under the precision-recall curves (PR-AUC) **b** on the held-out dataset are shown. **** $P < 0.0001$. P values are calculated using Wilcoxon test. **c** The features used by HRD_{CNA} to predict HRD and their contributions. **d** The difference in 10 CNA features between HRD and HRP samples. Q values are calculated using Wilcoxon test.

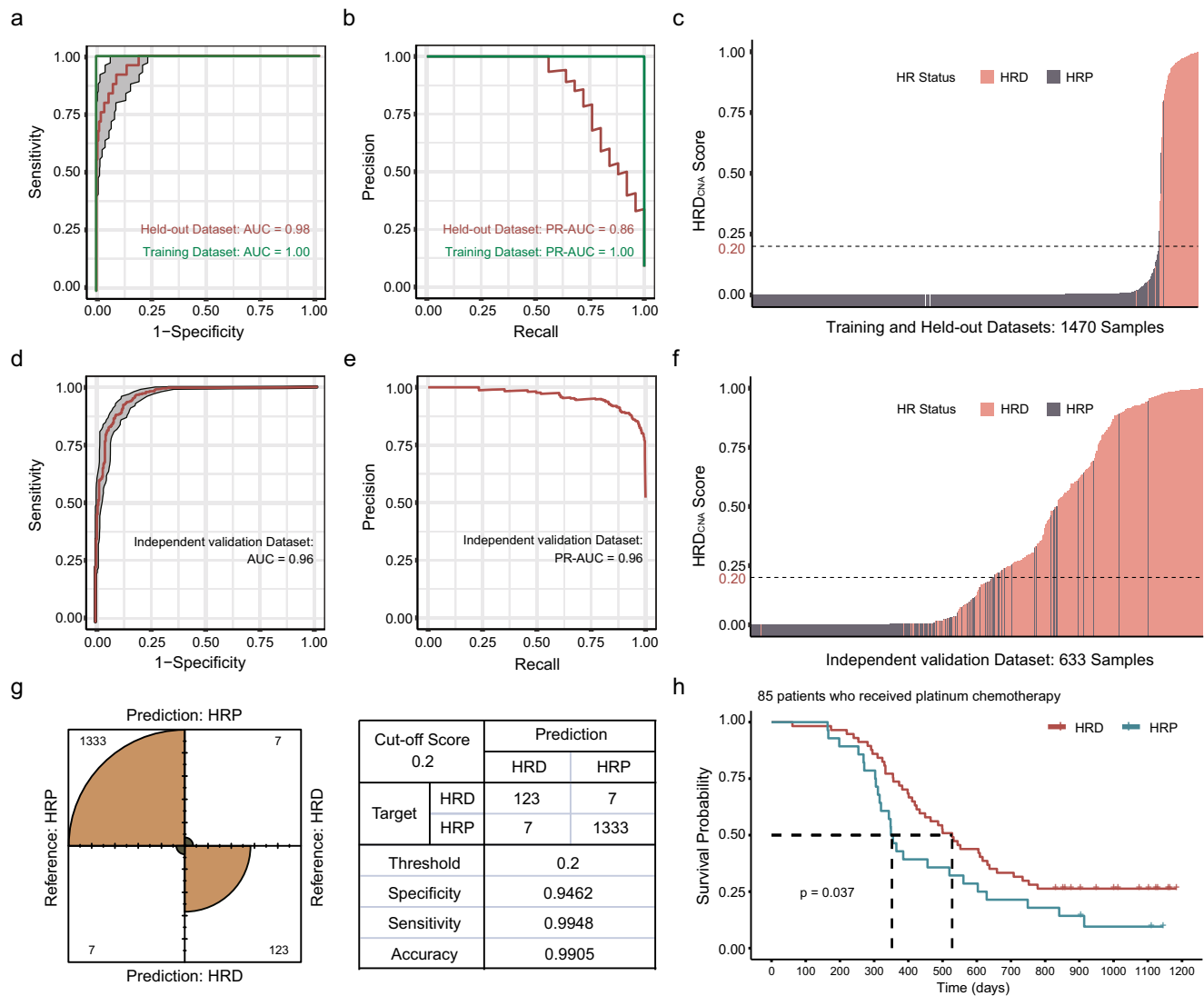


Fig. 3 Performance of HRD_{CNA} model. **a, b** ROC curves **a** and PR curves **b** analysis showed the performance of HRD_{CNA} built using CNA features on the training and held-out datasets. The gray shaded area represents a 95% confidence interval. **c** HRD_{CNA} scores for 1470 pan-cancer samples in training and held-out datasets are ordered from lowest to highest. The HRD_{CNA} scores of almost all HRD samples are above 0.2. **d, e** The ROC curve **d** and PR curve **e** showed the performance of HRD_{CNA} model in the independent validation dataset of 633 cancer samples. The gray shaded area represents a 95% confidence interval. **f** HRD_{CNA} scores for 633 pan-cancer samples in three independent validation datasets are ordered from lowest to highest. The horizontal dashed line shows a cut-off score of 0.2. **g** The performance of HRD_{CNA} model in training and held-out datasets is visualized with a confusion matrix under cut-off score of 0.2. **h** Kaplan-Meier (KM) survival analysis. Patients for analysis included 85 patients who had received platinum chemotherapy from 501 pan-cancer cohort. Samples are grouped based on HRD_{CNA} scores. Patients with high HRD_{CNA} scores showed significantly improved survival when treated with platinum chemotherapy. *P* value is calculated using log-rank test.

independent cohort of 71 TNBC patients (Supplementary Fig. 6a). The scatterplot displays the relationship between HRD_{CNA} score and HRDetect score (Supplementary Fig. 6b). Since HRD score is not applicable to panel sequencing data, we use data from WGS and SPN array sequencing to compare its performance with HRD_{CNA}, including 66 breast cancer WGS and SNP array samples. Compared with LST and HRD score, HRD_{CNA} has similar performance but is better than TAI and LOH (Supplementary Fig. 6c, d).

When evaluating the performance of the model, it is found that HRD_{CNA} scores of HRD samples are almost above 0.2, and HRP samples mostly get lower scores (Fig. 3c, f). Consequently, 0.2 was chosen as the threshold for HRD prediction (Fig. 3g). To further validate the performance of our model, we perform a Kaplan-Meier (KM) survival analysis on patients including 85 patients who had received platinum chemotherapy from 501 pan-cancer

cohort²⁴. Compared with patients with low HRD_{CNA} scores, patients with high HRD_{CNA} scores show significantly improved survival when treated with platinum chemotherapy (Fig. 3h). Then we compare our model with HRD score, a widely used HRD predicting method, and we define samples with HRD score ≥ 42 as HRD according to previous study²⁵ (Supplementary Fig. 7). Survival analysis suggests that HRD_{CNA} shows similar performance as HRD score.

BP10MB[1] and SS[>7 & <=8] as potential biomarkers of HRD. The counts of CNA features significantly differ between HRD and HRP samples (Fig. 2d). Among these 10 selected CNA features for HRD_{CNA} model construction, the top two important components are BP10MB[1] and SS[>7 & <=8], and either feature could predict HRD with comparable performance to the

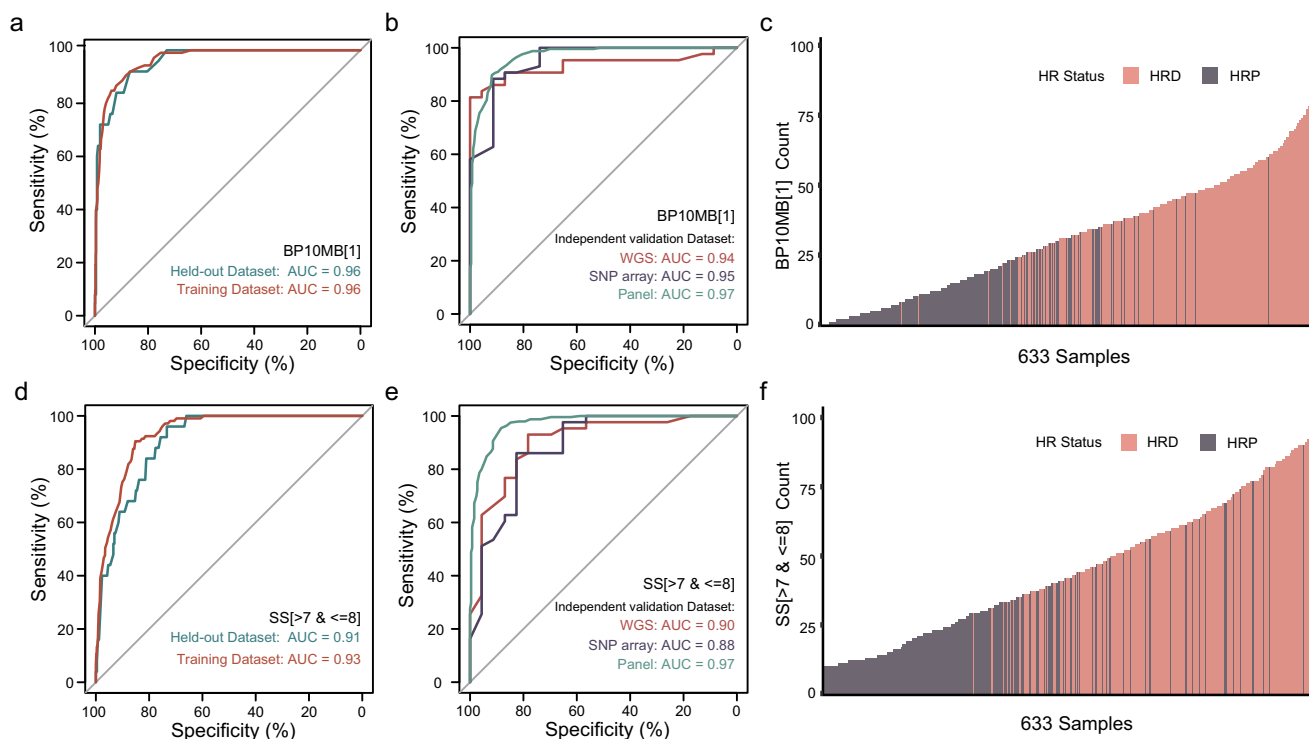


Fig. 4 BP10MB[1] and SS[>7 & <=8] as potential biomarkers of HRD. **a** HRD prediction ROC curves of CNA feature BP10MB[1] in training and held-out datasets. **b** ROC curves showed the performance of CNA feature BP10MB[1] in three independent validation datasets with different sequencing platforms, a total of 633 cancer samples, including 66 breast cancer samples with WGS sequencing, 66 breast cancer samples with SNP array sequencing, and 501 pan-cancer cancer samples with panel sequencing. **c** BP10MB[1] count for 633 cancer samples is ordered from lowest to highest. **d** ROC curves of CNA feature SS[>7 & <=8] in training and held-out datasets. **e** ROC curves showed the performance of CNA feature SS[>7 & <=8] in three independent validation datasets. **f** SS[>7 & <=8] count for 633 cancer samples is ordered from lowest to highest.

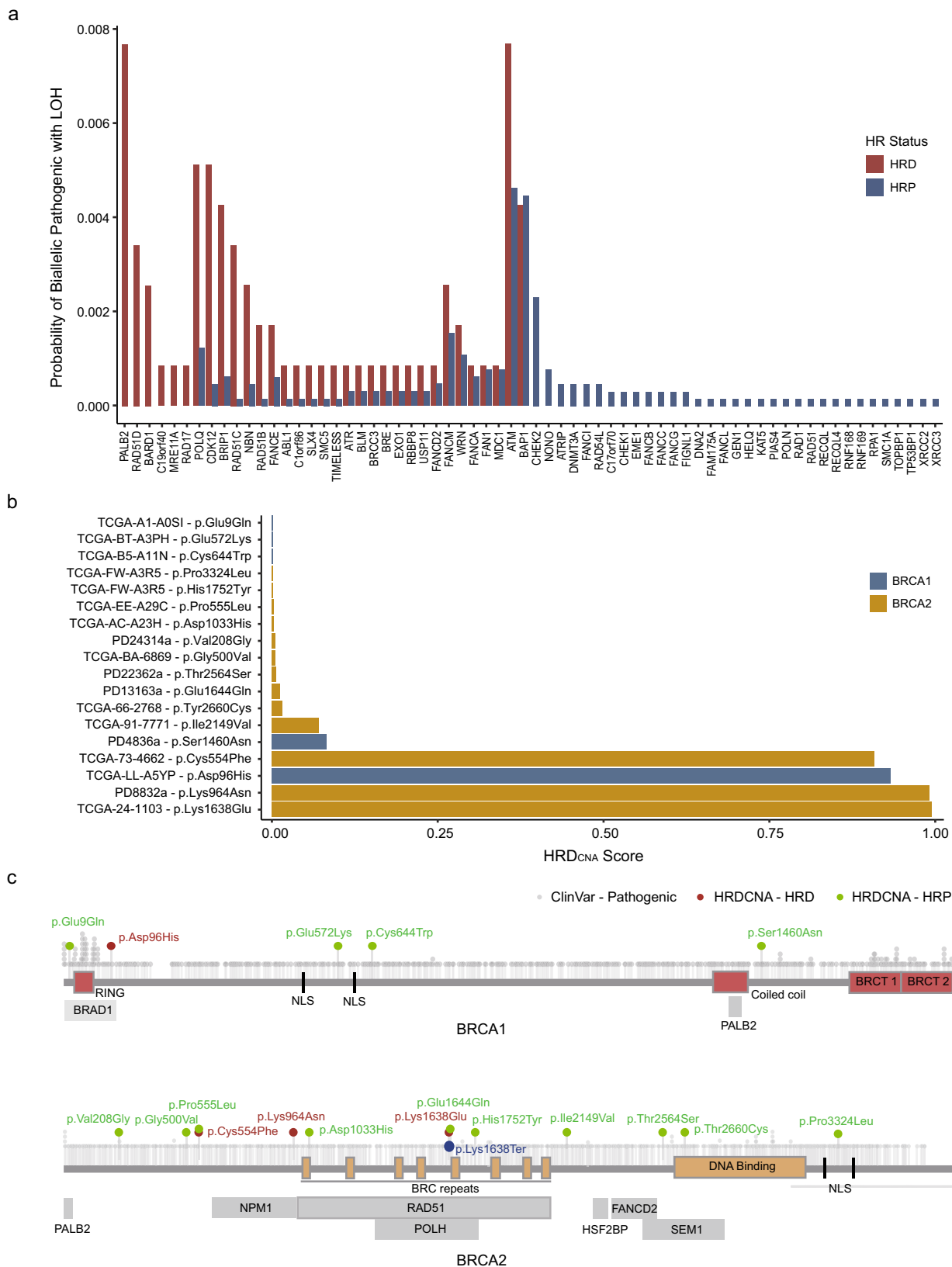
HRD_{CNA} model (Fig. 4a, d, Supplementary Fig. 8). BP10MB[1] indicates the number of breakpoints per 10MB of DNA is 1 and SS[>7 & <=8] indicates the log₁₀-based size of segments is greater than 7 and less than or equal to 8. The performance of BP10MB[1] and SS[>7 & <=8] is further verified with independent validation datasets (Fig. 4b, e). The counts of BP10MB[1] and SS[>7 & <=8] are enriched in HRD samples compared with HRP samples (Fig. 4c, f). Altogether, these results suggest that these two features BP10MB[1] and SS[>7 & <=8] may be potential biomarkers of HRD. Then we compare BP10MB[1]’s performance in HRD prediction with LOH, TAI, and LST in a cohort of 80 new breast cancers¹⁶ (Supplementary Fig. 9a). BP10MB[1] perform better than LOH and TAI, but similar to LST. LST, defined as chromosomal breaks between adjacent regions of at least 10 MB or larger, is also considered a robust indicator of HR status¹², the difference between LST and BP10MB[1] is shown (Supplementary Fig. 10). And BP10MB[1] also perform well in individual cancer types (Supplementary Fig. 9b).

Genetic basis for human HRD. BRCAness is functionally defined as a defect in homologous recombination repair that phenocopies loss of BRCA1/2²⁶. In addition to BRCA1/2, pathogenic variants in BRCAness genes such as partner and localizer of BRCA2 (*PALB2*) have been reported to contribute to human HRD^{9,27}. We explore potential BRCAness genes using the cancer genome atlas (TCGA) dataset by comparing the incidence of biallelic pathogenic mutations in HR-related genes²⁸ between patients predicted as HRD and HRP using HRD_{CNA} (Fig. 5a). The incidence of biallelic pathogenic mutations in *PALB2* and RAD51 paralog C (*RAD51C*) is observed to be significantly higher in HRD, and the inactivation of these two genes have been

demonstrated in previous studies as the major genetic basis of human HRD^{29–31}. Patients with biallelic inactivation mutations in the following genes: RAD51 paralog D (*RAD51D*) ($n = 4$), BRCA1-associated RING domain protein 1 (*BARD1*) ($n = 3$), Chromosome 19 open reading frame 40 (*C19orf40*) ($n = 1$) and MRE11 homolog A (*MRE11A*) ($n = 1$) are all predicted as HRD, suggesting these genes, especially *BARD1* and *RAD51D* inactivation can be genetic basis of human HRD.

HRD_{CNA} is applied to 33 cancers in the TCGA dataset, each sample gets an HRD_{CNA} score. The scores widely vary across different cancer types. The highest scores are observed in ovarian cancer, and many cancer types also have subsets of samples with high HRD_{CNA} scores. Some cancer types such as thyroid carcinoma, thymoma, uveal melanoma, and acute myeloid leukemia do not show high HRD_{CNA} scores (Supplementary Fig. 11). Similar pan-cancer HRD distribution patterns have been reported previously⁹.

BRCA VUS classification. The majority of missense mutations in BRCA1/2 have been classified as variants of uncertain significance (VUS) with unknown functional consequences. We calculate the HRD_{CNA} scores of 11,465 samples from 560 breast and TCGA datasets. Samples with BRCA VUS mutations and LOH in BRCA alleles are collected, and samples with known pathogenic HRD-causing genetic alterations are removed. HRD_{CNA} classified variants with scores above 0.2 as “likely pathogenic” and those with scores below 0.2 as “likely benign”. A total of 18 samples with homozygous BRCA VUS mutations are found, and 4 of them are classified as HRD based on HRD_{CNA} scores (Fig. 5b). Then we include the labeling information from other tools such as CancerVar, CADD, etc. for the ClinVar VUS clarified with HRD_{CNA} (Supplementary Table 1). Pathogenic BRCA1/2 missense



mutations in ClinVar are displayed (Fig. 5c), and functional domains are enriched with these pathogenic mutations^{32,33}. 5 variants with high HRDCNA scores are mainly distributed in the domains of BRCA1/2 proteins that interact with other proteins. For example, BRCA1 p.Asp96His is situated in the domain responsible for the interaction between BRCA1 and BARD1.

Discussion

Here we developed a pan-cancer HRD prediction tool HRDCNA based on CNA features. This tool enables accurate and robust HRD prediction using CNA profiles derived from a variety of cost-effective platforms, such as shallow WGS, SNP array, and panel sequencing. Furthermore, the CNA features BP10MB[1]

Fig. 5 Genetic basis for HRD and BRCA VUS reclassification. **a** The incidence of biallelic pathogenic mutations in HR-related genes in samples that are predicted to be HRD or HRP by HRD_{CNA} in the TCGA dataset. **b** The HRD_{CNA} scores for TCGA and 560 breast datasets of cancer patients with BRCA1/2 VUS and LOH are shown. **c** Distributions of VUS on BRCA proteins. BRCA1/2 amino acid loci of known pathogenic variants and pathogenic variants predicted by HRD_{CNA} are shown. The position of functional regions of BRCA 1/2 proteins and interacting proteins are obtained from UniPort databases, and pathogenic variants are defined based on ClinVar database. Vertical positions represent different kinds of amino acids. BARD1 BRCA1-associated RING domain protein 1, NLS nuclear localization sequence, PALB2 partner and localizer of BRCA2, BRCT BRCA1 C-terminus, BRC repeats ~30–40 residue long sequence regions in BRC2 protein, NPM1 nucleophosmin 1, RAD51 DNA repair protein RAD51 homolog 1, POLH DNA Polymerase Eta, HSF2BP Heat shock factor 2-binding protein, FACD2 Fanconi anemia group D2 protein, SEM1 26 S proteasome complex subunit SEM1.

and SS[>7 & <=8] are identified as the major contributors to HRD_{CNA} model, and also represent potential biomarkers for HRD. In addition to biallelic inactivation of *BRCA1* and *BRCA2*, *PALB2*, *RAD51C*, *RAD51D*, and *BARD1* inactivation are identified as the major genetic basis for human HRD. HRD_{CNA} tool also provides a new direction for effectively validating the pathogenicity of VUS. Carrying BRCA1/2 mutations, earlier stage, and lower levels of residual tumor after surgery have been proven as predictors of better prognosis for patients receiving platinum chemotherapy in ovarian cancer^{34,35}. Here, survival analysis shows that patients with high HRD_{CNA} scores have significantly improved survival than those with low HRD_{CNA} scores in patients receiving platinum chemotherapy. Over time, the application of HRD_{CNA} would unearth a substantial cohort of patients without BRCA1/2 mutations for chemotherapy using platinum drugs.

In addition to *BRCA1* and *BRCA2*, *PALB2* and *RAD51C* biallelic pathogenic mutations have been demonstrated in previous studies as the major genetic basis of human HRD^{29–31}. Notably, there is also a relatively high incidence of biallelic alterations in *BARD1* and *RAD51D* in HRD. BRCA1 partners with BARD1 to mediate the initial nucleotide excision of DNA damage and the recruitment of the recombinase RAD51³⁶. *BARD1* loss is sufficient to confer an HRD phenotype and significantly increase sensitivity to PARP inhibitors in prostate cancer cells³⁷. Triple-negative breast cancer cells carrying *RAD51D* K91fs and V200X variants are reported to be vulnerable to PARP inhibitors because they lead to decreased HR function³⁸. Together, our study suggests *BRCA1*, *BRCA2*, *PALB2*, *RAD51C*, *BARD1*, and *RAD51D* biallelic inactivation as the major genetic basis of human HRD, and it provides a novel opportunity for the precision medicine approaches in HRD and expands the population of patients who may benefit from agents targeting HRD.

Collectively, accurate measurement of HR status to find HRD patients has great clinical significance in precision medicine. HRD_{CNA} model is a CNA features-based robust tool for HRD detection. In the future, CNA features or signatures could be combined with artificial intelligence to reveal clinically meaningful CNA fingerprints for cancer precision diagnosis and therapy response prediction, and the HRD_{CNA} model developed in this study could represent one example of these clinically meaningful CNA fingerprints. HRD_{CNA} is freely available as an R package at <https://github.com/XSLiuLab/HRDCNA>.

Methods

Datasets. We have used CNA data from 1470 samples to develop the homologous recombination deficiency (HRD) gradient boosting machine (GBM) model, called HRD_{CNA} (Homologous recombination deficiency prediction by copy number alteration features), which included 1159 samples (WGS data) from PCAWG dataset⁹, 311 samples (SNP array data) from 560 breast dataset¹⁶. In particular, samples that were duplicated with PCAWG dataset in 560 breast dataset have been excluded, so there are no duplicated samples in our datasets. The information about the frequencies of cancer types in the training data for HRD_{CNA} model is shown in Supplementary Table 2. Somatic copy number data for the international cancer genome consortium (ICGC) portion of PCAWG dataset is downloaded at <https://dcc.icgc.org/releases/PCAWG/>. BRCA1/2 status annotations for this dataset are obtained from the supplementary data in Nguyen et al.⁹. Somatic copy number data of the 560 breast dataset are downloaded from the department of medical genetics at the University of Cambridge (<http://medgen.medschl.cam.ac.uk/serena->

<http://medgen.medschl.cam.ac.uk/serena-nik-zainal/>). BRCA1/2 status annotations and mutation data for this dataset are obtained from the supplementary data in Davies et al.¹⁶.

130 samples carrying mutations in BRCA1/2 are detected with loss of wild-type allele and labeled as HRD samples, and 1340 BRCA1/2 proficiency samples as HRP samples. A sample with one of the following events in BRCA1/2 is defined as a BRCA1/2 deficient sample: (i) complete copy number loss, (ii) LOH in combination with a pathogenic germline or somatic SNV/indel or structural variations, or (iii) pathogenic SNV/indels or structural variations in both alleles⁹. We select 80% of 1470 samples by random sampling as the training dataset, and 20% as the held-out dataset for model training. Copy number data, HR status information, and survival data for the panel dataset are obtained from the supplementary data in Wen H et al.²⁴. 66 breast dataset is publicly available in the figshare repository provided by de Luca et al.³⁹. Copy number data from ASCAT (array data) are available from <https://doi.org/10.6084/m9.figshare.9808496> and ascatsNGS (for original (<https://doi.org/10.6084/m9.figshare.9808505>) and downsampled WGS data are available from (30×: <https://doi.org/10.6084/m9.figshare.9808511>, 15×: <https://doi.org/10.6084/m9.figshare.9808514>, 10×: <https://doi.org/10.6084/m9.figshare.9808517>)).

Somatic copy number data of TCGA pan-cancer dataset are downloaded from genomic data commons data portal (<https://portal.gdc.cancer.gov/>). Allele-specific copy number analysis of tumors is performed using ASCAT2, to generate integral allele-specific copy number profiles for the tumor cells. Biallelic inactivation in HR-related genes information for TCGA dataset is obtained from Riaz et al.²⁸. The mutation data for TCGA dataset are obtained from the GDC Data Portal and Legacy Archive (<https://gdc.cancer.gov/>). The information on the cohort of 71 TNBC patients in TCGA dataset is derived from the supplementary data in Liao et al.⁴⁰. The information on the cohort of 80 breast patients is derived from the supplementary data in Davies et al.¹⁶. Pathogenicity annotations are obtained from ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>), CancerVar⁴¹ (<https://cancervar.wglab.org/>), CADD⁴² (<https://cadd.gs.washington.edu/>), MutationAssessor⁴³ (<http://mutationassessor.org/r3/>) and PolyPhen-2⁴⁴ (<http://genetics.bwh.harvard.edu/pph2/>).

Downsampling of WGS data. Downsampling of the original normal and tumor BAM files is performed using the samtools (version 1.3)⁴⁵ library function: samtools view -h -s x, where x represents the desired percentage of downsampling required (The values we use that will result in an approximate coverage of 30×, 15×, and 10×, depending on the coverage of the original data).

Construction of machine learning models. To find the best performance model among a multitude of methods, 9 machine learning models including extremely randomized trees (Extra trees), random forest, logistic regression, support vector machine (SVM), eXtreme gradient boosting (XGB), adaptive boosting (AdaBoost), decision tree, K-nearest neighbor (K-Neighbor) and gradient boosting machine (GBM) are trained, the AUC of the held-out dataset is selected as the performance criterion. All the hyper-parameters are determined by the gradient search method provided by the *GridSearchCV* function in Python package *sklearn*, and we use the default 5-fold CV strategy. The GBM model has the best performance on the held-out dataset. Thus, GBM is chosen to predict the probability of HRD. Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak base models, usually decision tree. R package *gbm* is used to implementation of the GBM.

Quantification of CNA signatures and CNA features. Sig-CNS is identified using R package *Sigminer*²⁰, which is based on the tool *SigProfiler*⁴⁶ caller. We pick up 8 signatures for our dataset due to its relatively high stability and low distance (Supplementary Fig. 12). Identification of Sig-CX used R package *CINSignatureQuantification*¹⁹. Calling CNA features is performed using R package *Sigminer*²⁰. 8 fundamental CNA features are computed, including the breakpoint count per 10 Mb (named BP10MB); the breakpoint count per chromosome arm (named BPArm); the copy number of the segments (named CN); the difference in copy number between adjacent segments (named CNCP); the lengths of oscillating copy number segment chains (named OsCN); the log₁₀ based size of segments (named SS); the minimal number of chromosome with 50% copy number variation (named NC50); the burden of chromosome (named BoChr). Then we classified 8 CNA feature distributions into 80 different components²⁰, and each copy number component has a clear biological meaning, for example, BP10MB[1] indicates the

number of breakpoints per 10MB of DNA is 1 (Details in Supplementary Table 3). Three models are built using CNA features directly or using two CNA signatures, Sig-CNS and Sig-CX, each of which is modeled repeatedly using Monte Carlo CV, which is repeated 100 times on the 80% dataset. The difference in AUC and PR-AUC of the three models is compared on the held-out dataset. CNA features are finally chosen for model development.

HRD_{CNA} training procedure. The model training includes a 4-step training procedure to obtain the final model (named HRD_{CNA}). Firstly, the HRD_{CNA} model is developed based on 76 significantly different CNA features among 80 CNA features (Supplementary Fig. 13) and the correlation matrix shows that no significant correlation is observed between most features (Supplementary Fig. 14). We choose bagging fraction 0.8 for more robust results. In addition, not to miss the optimal solution, we use a small learning rate of 0.01 together with a large number of trees (6000). We carry out 10-fold CV on the 80% training dataset and determine the best number of trees (572) from the Bernoulli deviance (Supplementary Fig. 15a). Then, from the relative influence score of features for the model, some CNA features are fewer contributions to our model, so we reduce model features. To calculate the contribution more accurately, we perform Monte Carlo CV on the training dataset using 76 CNA features, which are repeated 500 times. To calculate the relative influence of the specific variable, at each split in each tree, GBM computes the improvement in squared error, then averages the improvement made by each variable across all the trees that the variable is used²². We choose 10 CNA features with the top 10 relative influence scores among all features. Moreover, we again perform 10-fold CV on the training dataset using these 10 CNA features and determine the best number of trees (777) from the Bernoulli deviance (Supplementary Fig. 15b). Finally, we train our final model using 10 significantly different and important CNA features on the 80% training dataset.

Statistic and reproducibility. Data between two groups are compared using Wilcoxon rank-sum test (also known as “Mann–Whitney” test) depending on the normality of data distribution. In the survival analysis, the differences between different survival curves are compared using log-rank test, and the cut-off score is 0.14, which is determined based on the youden index⁴⁷. All reported *P*-values are two-tailed, and for all analyses, *P* ≤ 0.05 is considered statistically significant unless otherwise specified. All statistical analysis is performed using R v4.1.0. The area under the curves of the receiver operating characteristic curves (AUC) and the area under the precision–recall curves (PR-AUC) are then calculated by R package *precrec*.

CNA data to develop HRD_{CNA} includes 1470 samples with 1186 in the training dataset and 284 in the held-out dataset. 130 samples are labeled as HRD samples, and 1340 samples are labeled as HRP samples. The independent validation dataset consists of 633 cancer samples, including 66 breast cancer samples with WGS, 66 breast cancer samples with SNP array sequencing, and 501 pan-cancer samples with panel sequencing, and 330 of which are HRD samples and 303 HRP samples. TCGA dataset of 33 cancer types includes 10,906 samples. The cohort of 71 TNBC patients in TCGA dataset consists of 71 TNBC samples, 43 of which are HRD samples and 28 HRP samples. The cohort of 80 breast patients consists of 80 breast samples, 7 of which are HRD samples and 73 HRP samples.

Reporting summary. Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data used in this study are obtained from publicly available sources and are described in detail in Supplementary Table 4. The datasets generated and analyzed during the current study are available at <https://github.com/XSLiuLab/InterpretationAnalysisHRDCNA>⁴⁸. Any remaining information can be obtained from the corresponding author upon reasonable request.

Code availability

HRD_{CNA} is freely available as an R package at <https://github.com/XSLiuLab/HRDCNA>. The code used for data processing and generating the figures to reproduce the main results of this work is also publicly accessible at <https://github.com/XSLiuLab/InterpretationAnalysisHRDCNA>⁴⁸.

Received: 22 November 2022; Accepted: 2 May 2023;

Published online: 16 May 2023

References

- Mirza, M. R. et al. Niraparib maintenance therapy in platinum-sensitive, recurrent ovarian cancer. *N. Engl. J. Med.* **375**, 2154–2164 (2016).
- Li, H. et al. PARP inhibitor resistance: the underlying mechanisms and clinical implications. *Mol. Cancer* **19**, 107 (2020).
- Stover, E. H., Fuh, K., Constantinopoulos, P. A., Matulonis, U. A. & Liu, J. F. Clinical assays for assessment of homologous recombination DNA repair deficiency. *Gynecologic Oncol.* **159**, 887–898 (2020).
- Hoppe, M. M., Sundar, R., Tan, D. S. P. & Jeyasekharan, A. D. Biomarkers for homologous recombination deficiency in cancer. *J. Natl Cancer Inst.* **110**, 704–713 (2018).
- Esteller, M. et al. Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. *J. Natl Cancer Inst.* **92**, 564–569 (2000).
- Hsu, N. C. et al. Methylation of BRCA1 promoter region is associated with unfavorable prognosis in women with early-stage breast cancer. *PLoS ONE* **8**, e56256 (2013).
- Baldwin, R. L. et al. BRCA1 promoter region hypermethylation in ovarian carcinoma: a population-based study. *Cancer Res.* **60**, 5329–5333 (2000).
- Castroviejo-Bermejo, M. et al. A RAD51 assay feasible in routine tumor samples calls PARP inhibitor response beyond BRCA mutation. *Embo Mol. Med.* **10**, e9172 (2018).
- Nguyen, L., Martens, J. W. M., Van Hoeck, A. & Cuppen, E. Pan-cancer landscape of homologous recombination deficiency. *Nat. Commun.* **11**, 5584 (2020).
- Birkbak, N. J. et al. Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discov.* **2**, 366–375 (2012).
- Abkevich, V. et al. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br. J. Cancer* **107**, 1776–1782 (2012).
- Popova, T. et al. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res.* **72**, 5454–5462 (2012).
- Stewart, M. D. et al. Homologous recombination deficiency: concepts, definitions, and assays. *Oncologist* **27**, 167–174 (2022).
- Diossy, M. et al. Breast cancer brain metastases show increased levels of genomic aberration-based homologous recombination deficiency scores relative to their corresponding primary tumors. *Ann. Oncol.* **30**, 1406–1406 (2019).
- Zhu, S. et al. Homologous recombination deficiency (HRD) score in aggressive prostatic adenocarcinoma with or without intraductal carcinoma of the prostate (IDC-P). *Bmc Med.* **20**, 237 (2022).
- Davies, H. et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517 (2017).
- Sztupinski, Z. et al. Comparative assessment of diagnostic homologous recombination deficiency-associated mutational signatures in ovarian cancer. *Clin. Cancer Res.* **27**, 5681–5687 (2021).
- Steele, C. D. et al. Signatures of copy number alterations in human cancer. *Nature* **606**, 984–991 (2022).
- Drews, R. M. et al. A pan-cancer compendium of chromosomal instability. *Nature* **606**, 976–983 (2022).
- Wang, S. et al. Copy number signature analysis tool and its application in prostate cancer reveals distinct mutational processes and clinical outcomes. *PLoS Genet.* **17**, e1009557 (2021).
- Tao, Z. et al. The repertoire of copy number alteration signatures in human cancer. *Brief. Bioinform.* **24**, bbad053 (2023).
- Friedman, J. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
- Eeckhoutte, A. et al. ShallowHRD: detection of homologous recombination deficiency from shallow whole genome sequencing. *Bioinformatics* **36**, 3888–3889 (2020).
- Wen, H. et al. Homologous recombination deficiency in diverse cancer types and its correlation with platinum chemotherapy efficiency in ovarian cancer. *BMC Cancer* **22**, 550 (2022).
- Loibl, S. et al. Survival analysis of carboplatin added to an anthracycline/taxane-based neoadjuvant chemotherapy and HRD score as predictor of response-final results from GeparSixto. *Ann. Oncol.* **29**, 2341–2347 (2018).
- Byrum, A. K., Vindigni, A. & Mosammamaparast, N. Defining and modulating ‘BRCAness’. *Trends Cell Biol.* **29**, 740–751 (2019).
- Macchini, M. et al. Treatment opportunities and future perspectives for pancreatic cancer patients with germline BRCA1-2 pathogenic variants. *Cancer Treat. Rev.* **100**, 102262 (2021).
- Riaz, N. et al. Pan-cancer analysis of bi-allelic alterations in homologous recombination DNA repair genes. *Nat. Commun.* **8**, 857 (2017).
- Nones, K. et al. Whole-genome sequencing reveals clinically relevant insights into the aetiology of familial breast cancers. *Ann. Oncol.* **30**, 1071–1079 (2019).
- Staaf, J. et al. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nat. Med.* **25**, 1526–1533 (2019).
- Polak, P. et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* **49**, 1476–1486 (2017).

32. Clark, S. L., Rodriguez, A. M., Snyder, R. R., Hankins, G. D. & Boehning, D. Structure-function of the tumor suppressor BRCA1. *Comput Struct. Biotechnol. J.* **1**, e201204005 (2012).
33. Roy, R., Chun, J. & Powell, S. N. BRCA1 and BRCA2: important differences with common interests. *Nat. Rev. Cancer* **12**, 372 (2012).
34. Bolton, K. L. et al. Association between BRCA1 and BRCA2 mutations and survival in women with invasive epithelial ovarian cancer. *Jama* **307**, 382–390 (2012).
35. Chi, D. S. et al. Identification of prognostic factors in advanced epithelial ovarian carcinoma. *Gynecol. Oncol.* **82**, 532–537 (2001).
36. Tarsounas, M. & Sung, P. The antitumorigenic roles of BRCA1-BARD1 in DNA repair and replication. *Nat. Rev. Mol. Cell Biol.* **21**, 284–299 (2020).
37. Dillon, K. M. et al. PALB2 or BARD1 loss confers homologous recombination deficiency and PARP inhibitor sensitivity in prostate cancer. *npj Precis. Oncol.* **6**, 49 (2022).
38. Ma, D. et al. Molecular features and functional implications of germline variants in triple-negative breast cancer. *JNCI: J. Natl Cancer Inst.* **113**, 884–892 (2021).
39. de Luca, X. M. et al. Using whole-genome sequencing data to derive the homologous recombination deficiency scores. *Npj Breast Cancer* **6**, 33 (2020).
40. Liao, G. et al. Combined homologous recombination repair deficiency and immune activation analysis for predicting intensified responses of anthracycline, cyclophosphamide and taxane chemotherapy in triple-negative breast cancer. *BMC Med* **19**, 190 (2021).
41. Li, Q. et al. CancerVar: An artificial intelligence-empowered platform for clinical interpretation of somatic mutations in cancer. *Sci. Adv.* **8**, eabj1624 (2022).
42. Rentsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–d894 (2019).
43. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118 (2011).
44. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
45. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
46. Islam, S. M. A. et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics* **2**, 100179 (2022).
47. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32–35 (1950).
48. Yao, H. XSLiuLab/InterpretationAnalysisHRDCNA: An interpretation analysis of copy number alteration features in pan-cancer homologous recombination deficiency prediction and biology (github.com) (2023).

Acknowledgements

We thank ShanghaiTech University High Performance Computing Public Service Platform for computing services. We thank multi-omics facility, molecular and cell biology

core facility of ShanghaiTech University for technical help. This work is supported by Shanghai Science and Technology Commission (21ZR1442400), National Natural Science Foundation of China (31771373), and startup funding from ShanghaiTech University.

Author contributions

H.Y., H.L., and J.W. collected the data, developed the HRD_{CNA} method, and performed the computational analysis. T.W., W.N., K.D., C.W., G.W., Z.T., X.Z., J.C., and X.S. participated in critical project discussions. X.S.L. conceptualized the idea, designed, supervised the study, and wrote the manuscript with input from H.Y. and H.L.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-023-04901-3>.

Correspondence and requests for materials should be addressed to Xue-Song Liu.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Anam Akhtar and Christina Karlsson Rosenthal.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023