frontiers | Frontiers in Genetics

TYPE Methods
PUBLISHED 28 September 2022
DOI 10.3389/fgene.2022.990244

OPEN ACCESS

EDITED BY
Ilias Georgakopoulos-Soares,
University of California, San Francisco,
United States

REVIEWED BY
Xing Chen,
China University of Mining and Technology, China
Pan Deng,
Microsoft Research Asia (China), China

*CORRESPONDENCE
Jiayin Wang,
wangjiayin@mail.xjtu.edu.cn
Shuqun Zhang,
shuqun_zhang1971@163.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Cancer Genetics and Oncogenomics, a section of the journal Frontiers in Genetics

RECEIVED 09 July 2022
ACCEPTED 09 September 2022
PUBLISHED 28 September 2022

CITATION
Wang X, Xu Y, Zhang Y, Wang S, Zhang X, Yi X, Zhang S and Wang J (2022), HRD-MILN: Accurately estimate tumor homologous recombination deficiency status from targeted panel sequencing data.
Front. Genet. 13:990244.
doi: 10.3389/fgene.2022.990244

HRD-MILN: Accurately estimate tumor homologous recombination deficiency status from targeted panel sequencing data

Xuwen Wang[1,2†], Ying Xu[1,2†], Yinbin Zhang[3†], Shenjie Wang[1,2], Xuanping Zhang[1,2], Xin Yi[4], Shuqun Zhang[3*] and Jiayin Wang[1,2*]

[1]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China, [2]Shaanxi Engineering Research Center of Medical and Health Big Data, Xi'an Jiaotong University, Xi'an, China, [3]Department of Oncology, The Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China, [4]Geneplus-Beijing Institute, Beijing, China

Homologous recombination deficiency (HRD) is a critical feature guiding drug and treatment selection, mainly for ovarian and breast cancers. As it cannot be directly observed, HRD status is estimated on a small set of genomic instability features from sequencing data. The existing methods often perform poorly when handling targeted panel sequencing data; however, the targeted panel is the most popular sequencing strategy in clinical practices. Thus, we proposed HRD-MILN to overcome the computational challenges from targeted panel sequencing. HRD-MILN incorporated a multi-instance learning framework to discover as many loss of heterozygosity (LOH) associated with HRD status to cluster as possible. Then the HRD score is obtained based on the association between the LOHs and the cluster in the sample to be estimated, and finally, the HRD status is estimated based on the score.

In comparison experiments on targeted panel sequencing data, the Precision of HRD-MILN could achieve 87%, significantly improved from 63% reported by the existing methods, where the highest margin of improvement reached 14%. It also presented advantages on whole exome sequencing data. Based on our best knowledge, HRD-MILN is the first practical tool for estimating HRD status from targeted panel sequencing data and could benefit clinical applications.

KEYWORDS

homologous recombination deficiency, targeted panel sequencing, multi-instance learning model, cancer genomics, sequencing data analysis

1 Introduction

Homologous recombination repair deficiency (HRD) usually refers to a state of Homologous Recombination Repair (HRR) dysfunction at the cellular level. HRD is a more stable molecular marker of malignancy (Moore et al., 2019), whose positive status is often found in various malignant tumors including ovarian, breast, and pancreatic ductal

cancers (Pellegrino et al., 2020). Clinical studies have shown that cancer patients with HRD-positive status present highly sensitive to platinum-based chemotherapy and poly (ADP-ribose) polymerase (PARP) inhibitors (Konstantinopoulos et al., 2015; Telli et al., 2016). Thus, estimating HRD status in breast/ovarian cancer patients can expand the benefit population and improve prognosis (González-Martín et al., 2019; Ray-Coquard et al., 2019; Fasching et al., 2021). The development of PARP inhibitors as high-grade serous carcinoma of the ovary, fallopian tube, or peritoneum (HGSC) therapy resulted from the observation that BRCA mutations significantly increased the *in vitro* susceptibility of cancer cells to PARP inhibition (Bryant et al., 2005; Farmer et al., 2005).

Unfortunately, estimating HRD status is a complicated computational problem. The initial idea detects the related germline variations (Hoppe et al., 2018) or somatic mutations (GSM) on BRCA1/2 genes (Bell et al., 2011). But later studies reported a lot of negative examples (Mirza et al., 2016; Coleman et al., 2017; Pujade-Lauraine et al., 2017). It is suggested that more markers should be considered in HRD estimation. Powered by genome sequencing, current methods estimating HRD status are all NGS data-based (Sztupinszki et al., 2018). There are state-of-the-art methods for whole-exome sequencing (WES) or whole-genome sequencing (WGS) data. At present, the popular clinical sequencing assays for HRD have four categories: HRR-related gene mutation assays (Ledermann et al., 2016; Sherill-Rofe et al., 2019), Genomic Instability Score (GIS) (Alexandrov et al., 2015), mutation signature (Alexandrov et al., 2013; Alexandrov et al., 2020), and HRD functional assays. The clinical validity of HRD functional assays has not been well confirmed (Miller et al., 2020), as each type of method has its limitations, especially when handling targeted panel sequencing data. In the Background section, we discuss the computational issues on targeted panel sequencing data in detail.

Here, we provide HRD-MILN, a novel machine learning-based approach for estimating HRD status. It accurately and efficiently captures the genomic features of LOH from targeted panel sequencing data. Since it is hard to model the unclear/non-significant associations between a LOH mutation on the genomic level and the HRD status on the patient level, we use a supervised learning information imprecise multi-instance learning (MIL) framework to solve the critical computational issue. Comparison experiments on real sequencing data validate the MIL model. Based on our best knowledge, HRD-MILN is the first practical tool for estimating HRD status from panel sequencing data and could benefit clinical applications.

2 Background

The initial biomarker for HRD is GSM on BRCA1/2 genes (Bell et al., 2011; Kanchi et al., 2014). It is soon reported insufficient because HRR involves dozens of known genes, and abnormalities in these genes may also contribute to the HRD phenotype (Lee and Kopetz, 2022). There is no clear evidence that HRD can also arise through GSM or methylation of a broader set of HRR-related genes or other as-yet-undefined mechanisms (Radhakrishnan et al., 2014). Furthermore, clinical studies showed that it as a biomarker for predicting PARPi or platinum responses in HGSC patients cannot currently be established (Swisher et al., 2009; Swisher et al., 2017; Bernards et al., 2018). Some scholars emphasize mutational signature (MS) as a novel biomarker for judging HRD (Davies et al., 2017; O'Kane et al., 2017). Another opinion suggests the somatic copy number variations (SCNVs) imply genomic scars (Miller et al., 2020), e.g., telomeric allelic imbalance (TAI) (Birkbak et al., 2012) and large-scale state transition (LST) (Popova et al., 2012). From the ARIEL studies of rucaparib, LOH status can be a biomarker of PARPi response (Mirza et al., 2016). Thus, HRD biomarkers have three categories: 1) GSM based, 2) copy number variation based, and 3) LOH based. As LOH is composed of mutations and copy number variations, LOH is considered the most potential efficient biomarker (Abkevich et al., 2012).

The existing methods for estimating HRD status are developed based on the different HRD biomarkers or combinations. Some approaches are based on GSM in HRR-related genes (including BRCA1/2). Although GSM in BRCA1/2 significantly increased the *in vitro* Sensitivity of cancer cells to PARP inhibition (Bryant et al., 2005; Farmer et al., 2005), it is not sufficient. Furthermore, GSM in other HRR-related genes is associated with distinct sensitivities to PARPi (Marshall et al., 2019). Some approaches are based on the MS. This strategy analyzes MS mainly relies on mutational features, transcriptional strand bias, genomic distribution, and association analysis with genomic features to cluster and transform each type of mutation into a visual pattern. This type of approach has achieved good results in estimating cancers with HRD. However, it needs as much genome-wide information as possible is likely to offer greater specificity and Sensitivity (Miller et al., 2020), e.g., MutationalPatterns (Blokzijl et al., 2018) and YAPSA (Hübschmann et al., 2021). So they might work for WES or WGS (Goldfeder et al., 2016) but not for targeted panel sequencing. Moreover, this strategy lacks clinical evidence to support the efficacy prediction of PARP inhibitors, and its application is objectively limited by using paraffin-embedded samples for clinical testing. Some other approaches are based on genomic scar. There are two commercially available assays, the tumor BRCA mutation assays with an unweighted sum of GIS or the assessment of the sub-chromosomal LOH portion of the genome (Telli et al., 2016). For GIS, BRCA mutation-positive or GIS score ≥42 can be considered HRD-positive (Telli et al., 2016). The LOH test's predefined cut-off of 14% or more defines LOH-high. It is deemed to be positive for HRD (Bell et al., 2011). The utility of LOH or GIS showed good clinical validity in their ability to determine the BRCAwt subgroups that benefited more from

PARPi in the relapse platinum-sensitive setting (Miller et al., 2020). However, the accuracy of existing strategies of LOH or GIS is based primarily on the accuracy of the SCNV assay by the number of genomic scars for patient tumor samples. The current SCNV detection tools cannot accurately detect genomic scars. These false-positive genomic scars can misclassify the sample as false-positive for HRD.

Moreover, there were discrepant results in the HRD scores in different races, cancer species, and lifestyles or living conditions (Pellegrino et al., 2020). Thus, estimating the HRD status by a uniform threshold is problematic, an unweighted sum of GIS. Most importantly, such methods generally require high DNA loading, sequencing data volume, GSM covering the HRR signaling pathway, etc. The genomic distribution of target regions is often sparse and uneven on targeted panel sequencing data (Li et al., 2012; Talevich et al., 2016). Therefore, TAI and LST cannot be obtained, which may lead to low GIS on the panel data, thus misclassifying the sample to be tested as negative for HRD.

Some machine learning (ML) based approaches use HRD biomarkers to build ML models for estimating HRD status (Watkins et al., 2014; Chao et al., 2018; Nguyen et al., 2020). For example, HRDetect (Davies et al., 2017) used a lasso logistic regression model to identify six distinguishing MS predictive of BRCA1/BRCA2 deficiency (Gulhan et al., 2019). Using a machine learning approach instead of a single metric threshold approach has more significant advantages. It can effectively solve the accuracy problems and lack of generalization of the traditional HRD score calculation method. However, these methods have two disadvantages. 1) The premise of using supervised learning algorithms is that we have access to the labels of the training instances. However, we do not know the intrinsic connection between HRD biomarkers and HRD. These markers are only a manifestation of HRD, i.e., we cannot get the label of genomic scars to determine HRD (Watkins et al., 2014; Telli et al., 2016). 2) Due to the limitation of targeted panel sequencing, only LOH can be obtained, and obtaining other genomic scars information is difficult. Therefore, most methods are developed for WGS or WES but not targeted panel sequencing data.

3 Materials and methods

We proposed a new method to estimate HRD status based on a multiple instance learning framework. It is not reasonable that the existing ML model often adopts an aggressive strategy to obtain the training data (Davies et al., 2017; Miller et al., 2020): For an HRD-positive patient, assign all LOH (or LST, TAI) calls of this patient's positive labels from a medical view. Meanwhile, the false-positive genomic scars can also affect the accuracy of this strategy (Miller et al., 2020). The latest research now suggests that there must be an association between LOHs and assessment of HRD status (Miller et al., 2020), which means it is certain that the presence of one or more LOHs makes the sample positive for HRD, but precisely what that association is not yet clear. Our research has two main steps to estimate HRD status based on a multiple instance learning framework. One step is identifying the potential association pattern between the LOHs and HRD status during training. Another step is to calculate the HRD score based on the association between the LOHs and the clusters in the sample to be estimated. This way, we can estimate the HRD status without giving the LOH label.

3.1 Identifying the potential association pattern between the LOHs and HRD status

In our research, we can't get the label of LOH for estimating the HRD status. Therefore, we adopt the MIL (Maron and Lozano-P\'rez, 1997), which does not need category labels of instances, and the training package has category labels. Here, we set every sample as a package and each LOH status in every package as an instance. The core idea of the multi-instance learning method was that if one instance in the package were close enough to the calculated target concept point, it would be considered positive. However, due to the difference and complexity of the individual samples, the complexity and diversity of LOH, or the inaccuracy of the detection results, we modified multi-instance learning by proposing k target concept points (LOHs cluster) for detecting HRD. The input of HRD-MILN is a LOH file (TSV format), which is the result file of FACETS (Shen and Seshan, 2016) detection LOH. The output of our model is the HRD score, which is the prediction of the HRD status by HRD-MILN for a cancer sample. We collected 56-panel capture and 44 whole-exome sequencing samples to develop our model. All these samples are labeled with HRD status.

3.1.1 Features selection for LOH

The initial feature dimension of LOH is 14. However, redundant features may affect the performance of the models. Therefore, we also performed feature selection for HRD-MILN. Due to MIL being different from MI, we used two steps (Ablation studying and Calculating the importance of each feature) for feature selection for LOH. First, the number of maximum practical features is calculated by Ablation studying. Then, the valuable features of LOH are selected by Calculating the feature's importance.

3.1.2 Ablation studying

To the candidate the adequate number of features, we adopt the strategy of ablation experiment based on MILBoost (Viola et al., 2007). MILBoost is a feature selection method for MIL, which focuses on feature selection through the boosting framework. In our ablation experiments, we first fix random

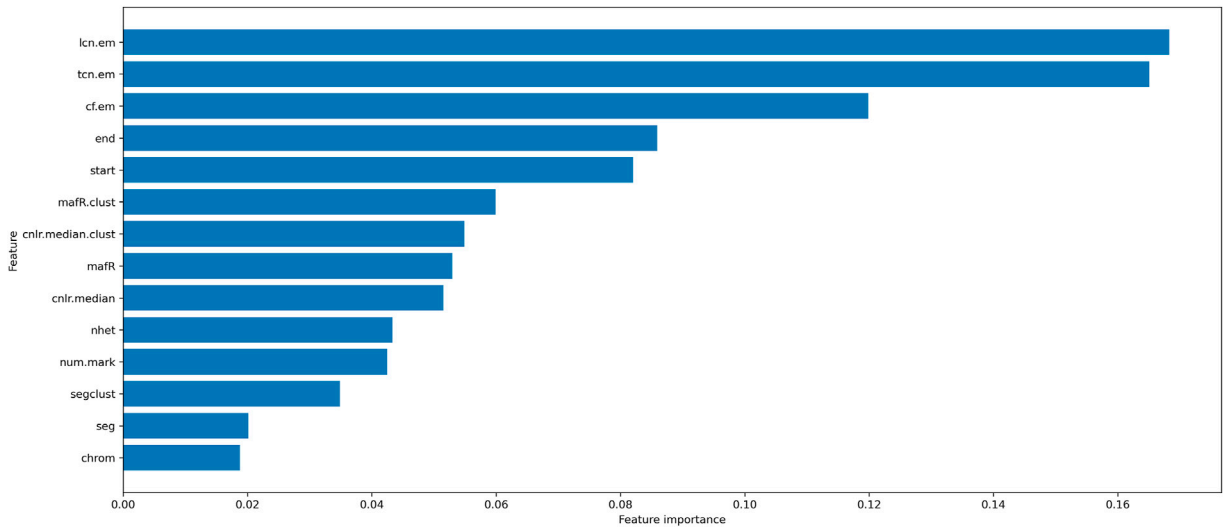

FIGURE 1
The importance of LOH features.

seeds and then observe and analyze the change in model performance as the number of features decreases. In each number-of-features experiment, i.e., when the number of features is fixed, we randomly select different features and perform 50 experiments, taking the mean value as a result. For each number of features, we served 100 experiments and finally took the mean value as the final result. We use the default parameters of MILBoost. Our experimental results show that the number of practical features of LOH is 9 (Supplemental Figure S1), so the default number of LOH features in HRD-MILN is 9. Of course, the reader can candidate the most effective number of LOHs according to their data type.

3.1.3 Calculating the importance of the feature

The initial features of LOH are (chrom, num. mark, nhet, cnlr. median, mafR, segcl, cnlr. median.clust, mafR.clust, start, end, cf. em, tcn. em, lcn. em). We used MILBoost to calculate each LOH feature's importance for developing the model (Figure 1). Combining the adequate number of features and Figure 1, we finally selected the following nine features, including (nhet, cnlr. median, mafR, mafR.clust, start, end, cf. em, tcn. em, and lcn. em) (Table 1), as well as the adequate number of features and the feature importance of LOH. On the other hand, as a machine learning framework, although we have prior knowledge of some features, these features of LOH may not directly imply HRD-positive susceptibility. Here, we analyzed these features according to the training data. We believe that more biological or medical research will explain the potential susceptibility in the future. Next, the min-max normalization is used on the feature attributes. We scaled the attribute data with a significant difference, which would fall into a small interval to improve the algorithm's convergence speed and detection accuracy.

3.1.4 The LOHs cluster for estimating HRD status

As shown in Figure 2, the LOH instance was regarded as a point, and one sample package had multiple LOH instances. The trajectories of these LOHs were treated as a manifold. For example, in point A, this intersection should satisfy every positive package passed through this point, and no negative sample package passed through it (It may be a target LOH). LOH instances and sample packages were subjected to a particular probability distribution. The diversity density (DD) (Maron and Lozano-P\'rez, 1997) function value of a LOH instance was this point's probability value, satisfying the potential positive or negative sample package distribution. One LOH instance had a DD value to find the max DD value as the target concept LOH. Then we used this target concept LOH as a reference to calculate the distance between every LOH and this LOH and then determine the HRD status of this sample through whether the minimum distance was within the threshold. Based on the complexity of our research content, we propose a strategy of multiple target concept LOHs and gather them into a cluster (named the LOHs cluster). Finally, the LOHs in the samples to be estimated are clustered with the LOHs cluster, and this process can filter the false-positive LOHs and LOHs that are not related to HRD status.

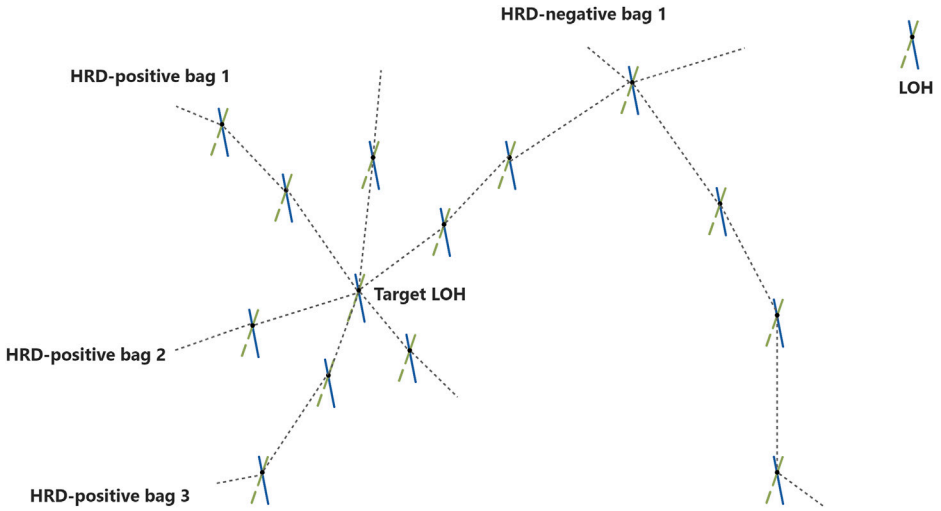

FIGURE 2
The schematic diagram of the target point. Showed in the figure, an instance was regarded as a point, and one package had multiple instances. The trajectories of these instances were treated as a manifold. For example, in point A, this intersection should satisfy that every positive package passed through this point, and no negative package passed through this point. So, point A may be a target point.

TABLE 1 The specific meaning of every feature attribution.

Feature	Specific meaning
chrom	The No. of chromosome
seg	ID number in this segment started from 1
num.mark	Detection intervals contained in this segment
nhet	Heterozygous SNP included in this segment
cnlr.median	The median of the copy number log ratio in this segment
segclust	This segment cluster was based on tcn and icn
cnlr.median.clust	The median of the copy number log ratio in this cluster
start	The start position
end	The end position
mafR	The summary statistic of log odd-ratio as described
mafR.clust	The summary statistic of log odd-ratio as described in this cluster
cf.em	The em value of the cell content in this segment
tcn.em	The em value of the total copy number in this segment
lcn.em	The em value of the less copy number in this segment

3.1.5 Identifying the association between the LOHs cluster and HRD status

The influence of every characteristic on the label could be modeled in the DD algorithm by associating an unknown factor. The target concept LOH, which means hot spot LOH, consisted of two values the ideal attribute value and the scale value. $T = \{t_1, t_2, \cdots, t_K\}$ represented the target concept LOHs, $t_k = \{t_{k1}, t_{k2}, \cdots, t_{km}\}$, t_{nd} represented d^{th} feature of t_k, m is the dimension of the feature, and $label(B_i|T)$ represented the prediction of B_i with T. B_{i+} represented the i^{th} HRD positive sample, B_{ij+} represented the j^{th} LOH instance in the i^{th} HRD positive sample. B_{ijd+} described the d_{th} feature of LOH in the j^{th} LOH instance of the i^{th} HRD positive sample. The same as the B_{i-}, B_{ij-}, B_{ijd-}.

$$Pr\left(t = t_k|B_{ij}\right) = \exp\left\{ -\sum_{d=1}^{m}\left(s_d\left(B_{ijd} - t_{kd}\right)\right)^2 \right\} \quad (1)$$

Pr is denoted as the probability of the LOH becoming a potential target concept point, defined as the distance between the LOH and the target concept point. The similarity between

the conformation and the ideal shape increased, and the bending strength decreased exponentially. Eq. 1 was used to determine if T is a real target concept of LOH. Since some of the features may be uncorrelated or need a higher weight, thus we used the weighted Euclidean distance (Ungerboeck, 2006) to measure the distance between LOHs, and s_d represented the weight of the d^{th} feature. The initial default value of s_d is 1. Readers can set it according to their own data.

3.2 Efficiently solving the model by modifying the EMDD algorithm

A single target concept LOH was used in the traditional experiment using the multi-instance learning method (Zhang and Goldman, 2001) algorithm (e.g., EMDD) to detect HRD status. However, the prediction of the category of the package would bed unsatisfactory. EMDD is a very typical and widely used multi-instance learning method. Thus, we choose EMDD as the research object and improve it to verify the effectiveness of our proposed multiple target concept points in detecting HRD. Our model processes a standard TSV file containing LOH and outputs a regular TSV file. To ensure that the relationship between K target LOHs is low, we used the weighted Euclidean distance to measure the distance among points. So, our strategy is suitable to solve the hot spots and randomness problems of LOH for HRD detection. The main steps are as follows.

E-step: We selected several initial LOH instances that were most likely to be labeled, different from the traditional EMDD algorithm. Then we used the current hypothetical target t to estimate the most probable label LOH in each training package (here means sample), and these LOH instances represented their respective packages. The *Pro* was used as the threshold of reliable candidate target LOH instances.

$$label\left(B_i|T\right) = \frac{\sum_{i=k}^{K} p_k^*}{K}, p_k^* > Pro, p_k^* \in Pr\left(B_{ij}|t = t_k\right) \quad (2)$$

$$NNLDD\left(T, D\right) = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n}\left(-\log Pr\left(l_i|t_k', B_i\right)\right)}{K} \quad (3)$$

$$t' = \left\{t_1', t_2', \cdots, t_k'\right\}, \prod Pr\left(l_i|t_k'\right) > Pro, k \in K \quad (4)$$

Here, *Label(Bi|T)* or l_i represented the candidate label of *Bi*. p_i* meant the possibility of candidate label of *Bi* with the target point *i*. *K* represented the number of target points. *NNLDD* represented DD values of *K* target points. *t'* represented the new concept point.

M-step: According to the above Eq. 2 Eq. 3 Eq. 4, we used the gradient ascent method to obtain the *K* new concept points *t'* for these training examples. Then we used the *t'* to replace the *t* in the E-step. Repeat E-step and M-step until the difference between the adjacent *t* values converges.

3.3 Calculating the HRD score to estimate HRD status

By the above steps, we can get the LOHs in the sample to be estimated similar to the LOHs cluster. Then, according to these LOHs, calculate the HRD score. Briefly as below: Bi represented the i^{th} sample to be expected, B_{ij} represented the j^{th} LOH instance in the i^{th} sample to be predicted. B_{ijk} described the k_{th} feature of LOH in the j^{th} LOH instance of the i^{th} sample. Next, we can calculate the probability of it being LOH instance positive (the HRD score) by Eq. 6. Then we can determine whether the LOH to be tested is a LOH instance positive by Eq. 5, Eq. 7. As long as there was a LOH instance positive in a sample, the sample was labeled positive HRD status, otherwise negative HRD status.

$$f_{HRD-MILN}\left(B_i\right) = \begin{cases} +1, \exists f\left(B_{ij}\right) = +1 \\ -1, \exists f\left(B_{ij}\right) = -1 \end{cases}, \left(1 \le i \le n, 1 \le j \le n\right) \quad (5)$$

$$Pro\left(B_{ij}\right) = \frac{\sum_{k=1}^{K} exp\left\{-\sum_{d=1}^{n}\left(s_d\left(B_{ijd} - t_{kd}\right)\right)^2\right\}}{K} \quad (6)$$

$$f\left(B_{ij}\right) = \begin{cases} = +1, Pro\left(B_{ij}\right) \ge NNLDD_{thre}^* \\ = -1, Pro\left(B_{ij}\right) \le NNLDD_{thre}^* \end{cases} \quad (7)$$

Here, $f_{HRD\text{-}MILN}(Bi)$ means the prediction of the HRD status of HRD-MILN. $NNLDD_{thre}$* means the sample's probability threshold is HRD positive, trained by E-step and M-step.

3.4 Data collection and bioinformatics pipeline

3.4.1 Targeted panel sequencing samples

We had panel capture sequencing data of 56 cancer samples with known HRD status, including 28 HRD positive and 28 HRD negative samples. And these samples were provided by Gene+, Inc. We worked with the BAM files, which were obtained from https://db.cngb.org/.

3.4.2 Whole-exome sequencing samples

3.4.2.1 Study design and patients

In the meantime, subjects recruited for this study included a subset of clinically diagnosed breast cancer (25 individuals) and ovarian cancer (19 individuals) patients in the Department of Oncology, the Second Affiliated Hospital of Xi'an Jiaotong University (Approval No. 2022038). The institutional review and privacy boards reviewed this trial at all sites. All patients provided written informed consent.

3.4.2.2 WES and LOH analysis

This cohort's available tumor tissues from 44 patients underwent whole-exome sequencing (WES). Genomic DNA

TABLE 2 Results under different Pro thresholds for selecting target points. Abbreviations: *h*: hour. *Pro*: the threshold of reliable candidate target points.

Pro	0.85	0.90	0.95	0.97
Average Targets	4	3	2	1
Average Time	2 h	1 h	0.5 h	0.3 h

was obtained from formalin-fixed, paraffin-embedded (FFPE), or aspirated biopsy tumor specimens and blood samples QIAamp DNA FFPE Tissue Kit and DNeasy Blood Tissue Kit (Qiagen, United States), respectively, and analyzed using the dsDNA HS detection kit (ThermoFisher Scientific, United States). All samples were sequenced on an Illumina Hiseq4000 instrument using the 150 PE protocol (Illumina, United States). The quality control of FASTQ files is dealt with by Trimmomaticc (Bolger et al., 2014). Paired-end reads were then mapped to the human reference genome (hg19) using BWA-MEM (v.0.7.15) (Li and Durbin, 2009). Duplicate reads were marked by the MarkDuplicates tool in Picard. GATK3 was used to process the resulting BAM files to correct mapping and base quality score recalibration (Van der Auwera et al., 2013). We used ContEst (Broad Institute, contamination rate <0.02) to estimate Cross-sample (Cibulskis et al., 2011). We used Mutect (Cibulskis et al., 2013) and Scalpel (Fang et al., 2016) to call Somatic Single Nucleotide Variant and insertion/deletions. We used Snp-pileup (Shen and Seshan, 2016) to generate a CSV file containing SNV information on each chromosome from each dataset's Bam file (BamN and BamT). Then, we used Facets (Shen and Seshan, 2016) to generate a *_cncf. TSV file containing copy number variations from the results of the Snp-pileup.

4 Results

To evaluate the performance of HRD-MILN, we conduct experiments on real datasets, which contain panel capture sequencing data of 56 cancer samples with known HRD status (28 HRD positive samples and 28 HRD negative samples), and a subset of WES samples of clinically diagnosed breast cancer (25 individuals) and ovarian cancer (19 individuals) patients. First, we did multiple experiments to demonstrate that multiple target LOHs affect the algorithm's accuracy in detecting HRD. We also did several experiments to compare the proposed method and the original algorithm (EMDD). Finally, we also did several experiments to compare the performance of detecting HRD between HRD-MILN and the existing algorithm (SigMA) (Gulhan et al., 2019). In addition, the performance of the above methods is quantified by Precision, Sensitivity, and f1-score, where $precision = TP/(TP + FP)$, $sensitivity = TP/(TP + FN)$, *and f1-score* is the harmonic mean between the *Precision* and *Sensitivity*. *TP* is the number of true positive HRD samples, FP denotes the number of false positive HRD samples, and *FN* represents the number of false negative HRD samples. Their default parameters are used to compare our method with existing ones fairly.

4.1 Application of HRD-MILN to targeted panel sequencing samples

Prediction of HRD status was a binary classification problem (Ledermann et al., 2016). Due to the small samples, the experiment used the 10-fold cross-validation method and Nested cross-validation. To select an appropriate target concept point and the appropriate Pro threshold, 500 sets of experiments were done. Through multiple sets of pre-experiments, we set the default Pro threshold of candidate target concept points to 0.9 (Table.2). Meanwhile, to verify the necessity of the HRD-MILN method to predict the HRD status based on the panel sequencing data, a total of 500 sets of experiments were also done. The variance of the 10-fold cross-validation is 0.00084. Nested cross-validation is very suitable for small-sample machine learning modeling. Varma et al. (Varma and Simon, 2006) show in their paper that the test set error obtained using nested cross-validation is almost the correct error. Comparing the scores of nested cross-validation with the regular procedure (Supplemental Figure S2) shows that the average difference is 0.000522 with std. dev. of 0.000920, it is again demonstrated that our proposed method is still valid in the case of small samples. The specific accuracy of the different targets model is shown in Table.3. From Table.3, it could be seen that the candidate target was 3, the average scores of Precision, Sensitivity, and f-score were all of the best, and the accuracy of each group was not fluctuate much. Note that the difference between our model and EMDD is that the number of target points is different (k vs. 1), and the number of the target point of EMDD is 1. According to the particularity of the HRD samples, it was necessary to improve the EMDD algorithm and propose constructing a multiple target concept point to assist decision-making in improving the accuracy of HRD detection. On the other hand, due to the particularity and complexity of HRD, introducing too many targets may introduce time complexity and background noise, which would affect the final experimental results.

Due to the difference and complexity of the individual samples, the complexity and diversity of LOH, or the inaccuracy of the detection results, it is hard for the EMDD algorithm (a single target concept in general) to detect HRD. Therefore, according to the characteristics of our study content, the EM was improved to help us better determine the HRD status of tumor samples more accurately. We compared the results tested by the two methods (Table.4; Figure 3). The improved HRD-MILN (the average Precision is

TABLE 3 Model accuracy in different targets on the panel sequencing data.

Groups	Average Targets	Precision	Sensitivity	F1-score
Group 1	1	0.25	0.5	0.33
	2	0.74	0.73	0.73
	3	0.89	0.88	0.88
	4	0.80	0.79	0.79
Group 2	1	0.76	0.54	0.63
	2	0.80	0.77	0.78
	3	0.87	0.86	0.86
	4	0.84	0.80	0.82
Group 3	1	0.76	0.54	0.63
	2	0.75	0.73	0.74
	3	0.85	0.85	0.85
	4	0.76	0.73	0.74
Group 4	1	0.52	1	0.68
	2	0.88	0.88	0.88
	3	0.91	0.91	0.91
	4	0.92	0.91	0.91
Group 5	1	0.82	0.71	0.76
	2	0.78	0.70	0.73
	3	0.79	0.75	0.77
	4	0.77	0.68	0.72
Group 6	1	0.76	0.54	0.63
	2	0.84	0.80	0.82
	3	0.90	0.89	0.89
	4	0.89	0.88	0.88
Group 7	1	0.76	0.54	0.63
	2	0.75	0.70	0.72
	3	0.89	0.89	0.89
	4	0.88	0.84	0.86
Group 8	1	0.81	0.80	0.80
	2	0.85	0.82	0.83
	3	0.88	0.87	0.87
	4	0.81	0.80	0.80
Group 9	1	0.90	0.88	0.89
	2	0.83	0.73	0.78
	3	0.91	0.90	0.90
	4	0.86	0.86	0.86
Group 10	1	0.83	0.75	0.79
	2	0.79	0.62	0.69
	3	0.86	0.85	0.85
	4	0.74	0.68	0.71

TABLE 4 HRD-MILN vs. EMDD. We compared various aspects of HRD detection performance of HRD-MILN and EMDD on the panel sequencing data. Abbreviations: HRD: Homologous recombination deficiency.

	Methods	Precision	Sensitivity	F1-score
Group 1	EMDD	0.25	0.5	0.33
	HRD-MILN	0.89	0.88	0.88
Group 2	EMDD	0.76	0.54	0.63
	HRD-MILN	0.87	0.86	0.86
Group 3	EMDD	0.76	0.54	0.63
	HRD-MILN	0.85	0.85	0.85
Group 4	EMDD	0.52	1	0.68
	HRD-MILN	0.91	0.91	0.91
Group 5	EMDD	0.82	0.71	0.76
	HRD-MILN	0.79	0.75	0.74
Group 6	EMDD	0.76	0.54	0.63
	HRD-MILN	0.90	0.89	0.89
Group 7	EMDD	0.76	0.54	0.63
	HRD-MILN	0.89	0.89	0.89
Group 8	EMDD	0.81	0.80	0.80
	HRD-MILN	0.88	0.87	0.87
Group 9	EMDD	0.90	0.88	0.88
	HRD-MILN	0.91	0.90	0.90
Group 10	EMDD	0.83	0.75	0.79
	HRD-MILN	0.86	0.85	0.85

0.88, Sensitivity is 0.87, F1-score is 0.87) is significantly better than EMDD (the average Precision is 0.72, Sensitivity is 0.68, F1-score is 0.70) in detecting HRD. To verify the effectiveness and necessity of the HRD-MILN method for panel sequencing data, we compared HRD-MILN with SigMA, the best model for detecting HRD based on panel sequencing data (Figure 3). HRD-MILN had higher scores (with a precision of 0.88, sensitivity 0.87, F1-score 0.87) than SigMA (with precision 0.76, sensitivity 0.63, F1-score 0.62) on each Evaluation indicators. This also proved the validity, accuracy, and necessity of HRD-MILN.

At the same time, we also compared HRD-MILN with ML for detecting HRD based on panel sequencing data (Figure 4). HRD-MILN had higher scores (with a Precision of 0.88, the Sensitivity of 0.87, F1-score 0.87) than machine learnings (MLS) (the best ML scores are precision 0.81, sensitivity 0.65, F1-score 0.72) on each Evaluation indicators. This

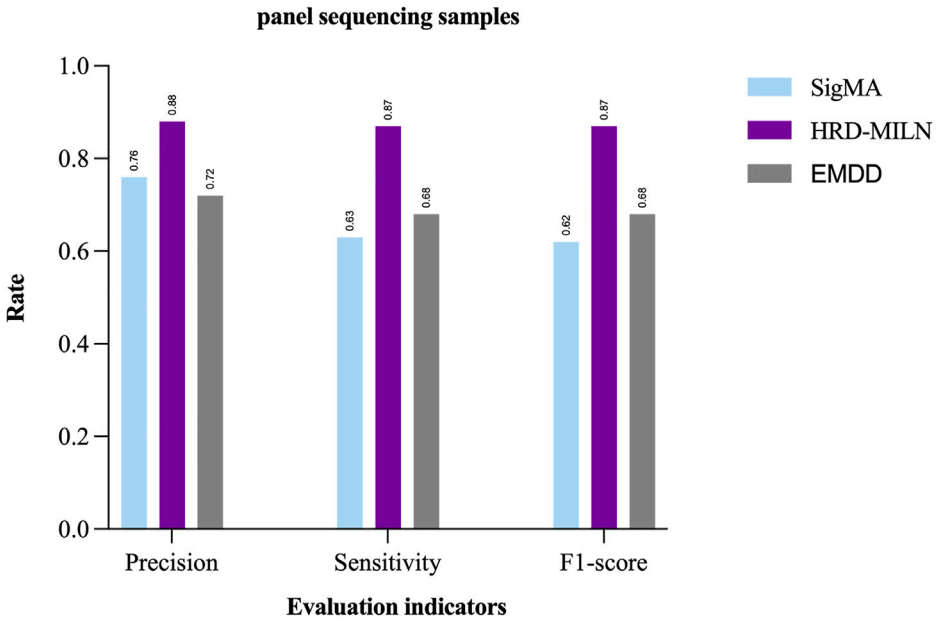

FIGURE 3
The performances of SigMA, EMDD, and HRD-MILN in panel sequencing samples. HRD-MILN had the highest scores (precision 0.88, sensitivity 0.87, F1-score 0.87) than SigMA and EMDD on each Evaluation indicator.

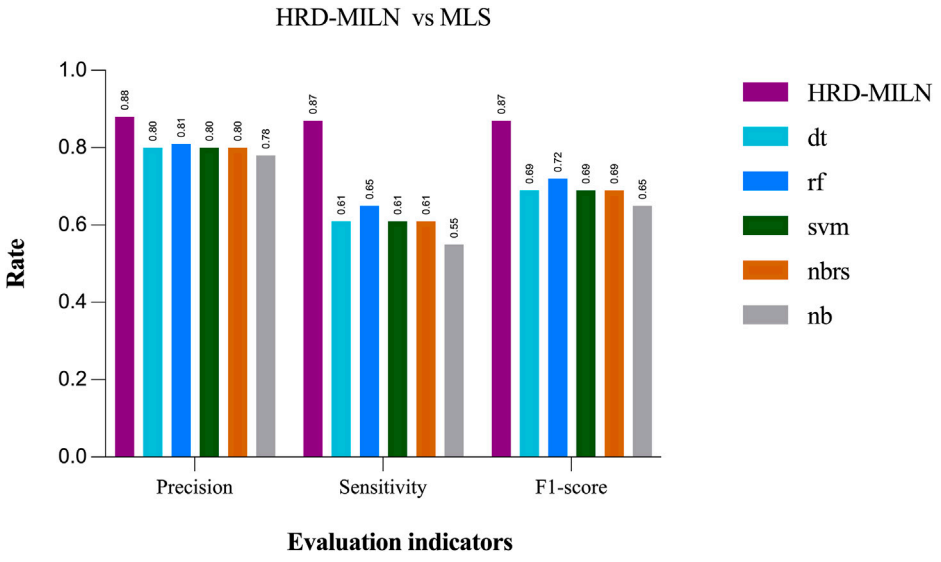

FIGURE 4
The performances of MILS and HRD-MILN in panel sequencing samples. HRD-MILN had the highest scores (precision 0.88, sensitivity 0.87, F1-score 0.87) than MILS on each Evaluation indicator. Abbreviations: MLS: machine learnings, dt: DecisionTree. rf: Random Forest. svm: Support Vector Machine. nbrs: KNeighbors. nb: GaussianNB.

TABLE 5 Model accuracy in different targets on the WES data.

Groups	Average Targets	Precision	Sensitivity	F1-score
Group 1	1	0.76	0.54	0.63
	2	0.62	0.62	0.62
	3	0.85	0.92	0.88
	4	0.63	0.57	0.6
Group 2	1	0.67	1	0.8
	2	0.7	0.73	0.71
	3	0.85	0.93	0.89
	4	0.73	0.86	0.79
Group 3	1	0.73	0.66	0.69
	2	0.67	0.73	0.7
	3	0.85	0.93	0.89
	4	0.72	0.96	0.82
Group 4	1	0.69	0.71	0.7
	2	0.81	0.87	0.84
	3	0.88	0.93	0.9
	4	0.84	0.86	0.85

also proves that ML is not suitable for detecting HRD. The MLS compared in our experiments are DecisionTree (dt) (Quinlan, 1986), Random Forest (rf) (Breiman, 2001), Support Vector Machine (svm) (Cho and Prabhu, 2002), KNeighbors (nbrs) (Blough et al., 2006), GaussianNB (nb) (De Moraes and Machado, 2008).

4.2 Application of HRD-MILN to whole-exome sequencing samples

To evaluate the effectiveness of our proposed method, we also chose to validate it on the WES sequencing samples. These samples included a subset of patients diagnosed with breast cancer (25 individuals) and ovarian cancer (19 individuals) in the Department of Oncology, the Second Affiliated Hospital of Xi'an Jiaotong University. For a fair comparison, we compared HRD-MILN with SigMA and EMDD by using the default parameters for each model.

First, we tested the effect of different numbers of target points on the performance of HRD-MILN. Through multiple sets of pre-experiments, we set the default Pro threshold of candidate target concept points to 0.9 (Table.2). Meanwhile, to verify the necessity of the HRD-MILN method to predict the HRD status based on the WES sequencing data by bootstrapping (sampling 500 random sets of real samples with replacement). Here we show only four sets of experimental results. The specific accuracy of the different targets model is shown in Table.5. From Table.5, it could be seen that the candidate target was 3. The average scores of Precision, Sensitivity, and f1-score were all of the best (0.86, 0.93, 0.89) compared with (0.73, 0.81, 0.77) for 4 targets, the second-best performer. The accuracy of each group did not fluctuate much. This result was the same as the panel

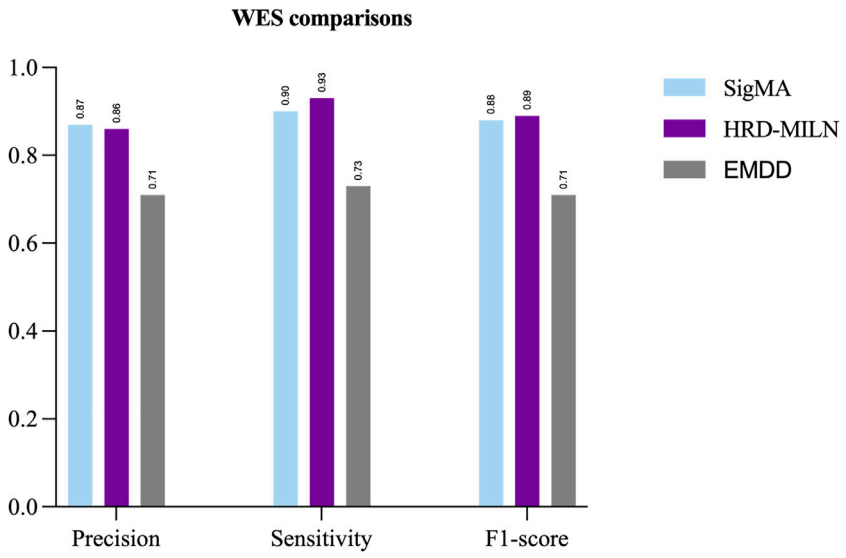

FIGURE 5
The performances of SigMA, EMDD, and HRD-MILN in whole-exome sequencing samples. HRD-MILN had the highest scores (precision 0.86, sensitivity 0.93, F1-score 0.89) than SigMA and EMDD on each Evaluation indicator.

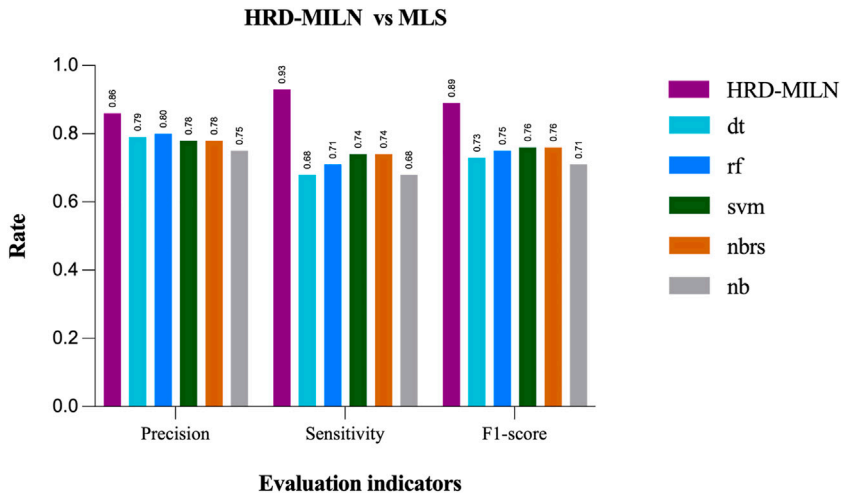

FIGURE 6
The performances of MILS and HRD-MILN in whole-exome sequencing samples. HRD-MILN had the highest scores (precision 0.86, sensitivity 0.93, F1-score 0.89) than MILS on each Evaluation indicator. Abbreviations: MLS: machine learnings, dt: DecisionTree. rf: Random Forest. svm: Support Vector Machine. nbrs: KNeighbors. nb: GaussianNB.

sequencing data. It again justifies our improvement strategy of using multiple target concept points to detect the status of HRD for cancer samples. Next, we compared the HRD-MILN with SigMA and EMDD on the WES sequencing samples. The results (Figure 5) show that HRD-MILN achieved the best average Precision of 0.86, Sensitivity of 0.93, and F1-score of 0.89 on the WES sequencing. SigMA is the second-best performer (0.87, 0.90, 0.88). From Figure 3, HRD-MILN was significantly better than SigMA and EMDD. Each method had relatively balanced scores on each Evaluation metric. The performance of EMDD is worse than HRD-MILN. This demonstrates that the improved approach is more effective for detecting HRD. Meanwhile, we compared HRD-MILN with ML for detecting HRD based on whole-exome sequencing data (Figure 6). HRD-MILN had higher scores than MLS (the best ML scores are precision 0.78, sensitivity 0.74, F1-score 0.76) on each Evaluation indicators. This also proves that ML is not suitable for detecting HRD on WES.

5 Discussion and conclusion

Accurately estimating HRD status is a challenging computational problem in cancer genomics and is also a bottleneck preventing from identifying potential benefits to patients. The mutational events on the HRR pathway and genomic scars (LOH, LST, TAI) suggest HRD estimation biomarkers. The existing ML model often adopts an aggressive strategy to obtain the training data: For an HRD-positive patient, assign all LOH (or LST, TAI) calls of this patient's positive labels. This strategy is not reasonable from a medical view. There are no significant associations between one LOH (or LST, TAI) at the genomic level and the HRD status at the patient level. Literature suggests that those biomarkers, many of which may be similar to passenger somatic mutations, may randomly occur on genomes. But another opinion considers those functional biomarkers identical because they are induced as genomic scars. Thus, the multi-instance learning framework seems the best solution at present to model the complicated associations/similarities. In this study, we incorporated multi-instance learning in a novel way. For the training instance, a LOH has complex genomic features. It also implies an individual difference. The complexity of LOH leads to the design of multiple target concept points. Thus, we selected more than one target concept point, which improved the accuracy and Sensitivity as expected. Thus HRD-MILN cloud solves the key computational issue that it is hard to model the unclear/non-significant associations between a LOH mutation on the genomic level and the HRD status on the patient level. And by establishing the intrinsic associations among HRD biomarkers and HRD status, HRD-MILN is much less sensitive to false positive mutation calls (e.g., LOHs) than the existing methods.

Targeted panel sequencing is the most popular sequencing strategy in clinical practices, not only because of the high cost-performance ratio but also due to governmental policies. It will keep the top cancer sequencing service providers over the coming years. Thus, it is meaningful to develop this tool for targeted panel sequencing data and hopefully could benefit cancer

patients. In addition, the proposed method could also be used on WES and WGS data. The experiments demonstrated that HRD-MILN consistently outperforms the existing methods on different sequencing data, which should be helpful for widespread clinical applications.

In the future, we will pursue two experimental aims. First, HRD clinical samples are challenging to collect. Although the number of HRD samples in this study is 'big data' compared to clinical studies, it is limited compared to model development. Therefore, we will continue to collect more HRD samples to verify the validity of the proposed method. Second, it would be worthwhile to investigate whether other clinical computing problems with HRD could benefit from HRD-MILN.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Ethics statement

The studies involving human participants were reviewed and approved by The medical ethics committee of Second Affiliated Hospital of Xi'an Jiaotong University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

JW and SZ conceived this research; XW, YX, SW, and XZ designed the algorithm and the method; XW implemented coding and designed the software; YZ and YX collected the patients and sequencing data; XW, YX, and SW performed the experiments and analyzed the data; XW and JW wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by Shaanxi's Natural Science Basic Research Program, grant number 2020JC-01. The APC was funded by Shaanxi's Natural Science Basic Research Program, grant number 2020JC-01.

Acknowledgments

We thank all faculty members and graduate students who discussed the mathematical and statistical issues in seminars.

Conflict of interest

Author XY is employed by Geneplus-Beijing Institute.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.990244/full#supplementary-material

References

Abkevich, V., Timms, K. M., Hennessy, B. T., Potter, J., Carey, M. S., Meyer, L. A., et al. (2012). Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br. J. Cancer* 107 (10), 1776–1782. doi:10.1038/bjc.2012.451

Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., et al. (2015). Clock-like mutational processes in human somatic cells. *Nat. Genet.* 47 (12), 1402–1407. doi:10.1038/ng.3441

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., et al. (2020). The repertoire of mutational signatures in human cancer. *Nature* 578 (7793), 94–101. doi:10.1038/s41586-020-1943-3

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., et al. (2013). Signatures of mutational processes in human cancer. *Nature* 500 (7463), 415–421. doi:10.1038/nature12477

Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D. W., Dao, F., et al. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* 474 (7353), 609–615. doi:10.1038/nature10166

Bernards, S. S., Pennington, K. P., Harrell, M. I., Agnew, K. J., Garcia, R. L., Norquist, B. M., et al. (2018). Clinical characteristics and outcomes of patients with BRCA1 or RAD51C methylated versus mutated ovarian carcinoma. *Gynecol. Oncol.* 148 (2), 281–285. doi:10.1016/j.ygyno.2017.12.004

Birkbak, N. J., Wang, Z. C., Kim, J.-Y., Eklund, A. C., Li, Q., Tian, R., et al. (2012). Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discov.* 2 (4), 366–375. doi:10.1158/2159-8290.CD-11-0206

Blokzijl, F., Janssen, R., van Boxtel, R., and Cuppen, E. (2018). MutationalPatterns: Comprehensive genome-wide analysis of mutational processes. *Genome Med.* 10 (1), 33. doi:10.1186/s13073-018-0539-0

Blough, D. M., Leoncini, M., Resta, G., and Santi, P. (2006). The k-neighbors approach to interference bounded and symmetric topology control in ad hoc networks. *IEEE Trans. Mob. Comput.* 5 (9), 1267–1282. doi:10.1109/TMC.2006.139

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30 (15), 2114–2120. doi:10.1093/bioinformatics/btu170

Breiman, L. (2001). Random forests. *Mach. Learn.* 45 (1), 5–32. doi:10.1023/A:1010933404324

Bryant, H. E., Schultz, N., Thomas, H. D., Parker, K. M., Flower, D., Lopez, E., et al. (2005). Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature* 434 (7035), 913–917. doi:10.1038/nature03443

Chao, A., Lai, C.-H., Wang, T.-H., Jung, S.-M., Lee, Y.-S., Chang, W.-Y., et al. (2018). Genomic scar signatures associated with homologous recombination deficiency predict adverse clinical outcomes in patients with ovarian clear cell carcinoma. *J. Mol. Med.* 96 (6), 527–536. doi:10.1007/s00109-018-1643-8

Cho, S., and Prabhu, V. V. (2002). A vector space model for variance reduction in single machine scheduling. *IIE Trans.* 34 (11), 933–952. doi:10.1023/A:1016126413117

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., et al. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* 31 (3), 213–219. doi:10.1038/nbt.2514

Cibulskis, K., McKenna, A., Fennell, T., Banks, E., DePristo, M., and Getz, G. (2011). ContEst: Estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* 27 (18), 2601–2602. doi:10.1093/bioinformatics/btr446

Coleman, R. L., Oza, A. M., Lorusso, D., Aghajanian, C., Oaknin, A., Dean, A., et al. (2017). Rucaparib maintenance treatment for recurrent ovarian carcinoma after response to platinum therapy (ARIEL3): A randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet* 390 (10106), 1949–1961. doi:10.1016/S0140-6736(17)32440-6

Davies, H., Glodzik, D., Morganella, S., Yates, L. R., Staaf, J., Zou, X., et al. (2017). HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* 23 (4), 517–525. doi:10.1038/nm.4292

De Moraes, R. M., and Machado, L. D. S. (2008). "Online training assessment in virtual reality simulators based on Gaussian naive bayes," in *Computational intelligence in decision and control* (World Scientific), 1147–1152.

Fang, H., Bergmann, E. A., Arora, K., Vacic, V., Zody, M. C., Iossifov, I., et al. (2016). Indel variant analysis of short-read sequencing data with Scalpel. *Nat. Protoc.* 11 (12), 2529–2548. doi:10.1038/nprot.2016.150

Farmer, H., McCabe, N., Lord, C. J., Tutt, A. N. J., Johnson, D. A., Richardson, T. B., et al. (2005). Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* 434 (7035), 917–921. doi:10.1038/nature03445

Fasching, P. A., Link, T., Hauke, J., Seither, F., Jackisch, C., Klare, P., et al. (2021). Neoadjuvant paclitaxel/olaparib in comparison to paclitaxel/carboplatinum in patients with HER2-negative breast cancer and homologous recombination deficiency (GeparOLA study). *Ann. Oncol.* 32 (1), 49–57. doi:10.1016/j.annonc.2020.10.471

Goldfeder, R. L., Priest, J. R., Zook, J. M., Grove, M. E., Waggott, D., Wheeler, M. T., et al. (2016). Medical implications of technical accuracy in genome sequencing. *Genome Med.* 8 (1), 24. doi:10.1186/s13073-016-0269-0

González-Martín, A., Pothuri, B., Vergote, I., DePont Christensen, R., Graybill, W., Mirza, M. R., et al. (2019). Niraparib in patients with newly diagnosed advanced ovarian cancer. *N. Engl. J. Med.* 381 (25), 2391–2402. doi:10.1056/NEJMoa1910962

Gulhan, D. C., Lee, J. J.-K., Melloni, G. E. M., Cortés-Ciriano, I., and Park, P. J. (2019). Detecting the mutational signature of homologous recombination deficiency in clinical samples. *Nat. Genet.* 51 (5), 912–919. doi:10.1038/s41588-019-0390-2

Hoppe, M. M., Sundar, R., Tan, D. S. P., and Jeyasekharan, A. D. (2018). Biomarkers for homologous recombination deficiency in cancer. *J. Natl. Cancer Inst.* 110 (7), 704–713. doi:10.1093/jnci/djy085

Hübschmann, D., Jopp-Saile, L., Andresen, C., Krämer, S., Gu, Z., Heilig, C. E., et al. (2021). Analysis of mutational signatures with yet another package for signature analysis. *Genes Chromosom. Cancer* 60 (5), 314–331. doi:10.1002/gcc.22918

Kanchi, K. L., Johnson, K. J., Lu, C., McLellan, M. D., Leiserson, M. D. M., Wendl, M. C., et al. (2014). Integrated analysis of germline and somatic variants in ovarian cancer. *Nat. Commun.* 5 (1), 3156. doi:10.1038/ncomms4156

Konstantinopoulos, P. A., Ceccaldi, R., Shapiro, G. I., and D'Andrea, A. D. (2015). Homologous recombination deficiency: Exploiting the fundamental vulnerability of ovarian cancer. *Cancer Discov.* 5 (11), 1137–1154. doi:10.1158/2159-8290.CD-15-0714

Ledermann, J. A., Drew, Y., and Kristeleit, R. S. (2016). Homologous recombination deficiency and ovarian cancer. *Eur. J. Cancer* 60, 49–58. doi:10.1016/j.ejca.2016.03.005

Lee, M. S., and Kopetz, S. (2022). Are homologous recombination deficiency mutations relevant in colorectal cancer? *J. Natl. Cancer Inst.* 114 (2), 176–178. doi:10.1093/jnci/djab170

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25 (14), 1754–1760. doi:10.1093/bioinformatics/btp324

Li, J., Lupat, R., Amarasinghe, K. C., Thompson, E. R., Doyle, M. A., Ryland, G. L., et al. (2012). Contra: Copy number analysis for targeted resequencing. *Bioinformatics* 28 (10), 1307–1313. doi:10.1093/bioinformatics/bts146

Maron, O., and Lozano-P\'rez, T. s. (1997). A framework for multiple-instance learning. *Nips* 97, 570–576.

Marshall, C. H., Sokolova, A. O., McNatty, A. L., Cheng, H. H., Eisenberger, M. A., Bryce, A. H., et al. (2019). Differential response to olaparib treatment among men with metastatic castration-resistant prostate cancer harboring BRCA1 or BRCA2 versus ATM mutations. *Eur. Urol.* 76 (4), 452–458. doi:10.1016/j.eururo.2019.02.002

Miller, R. E., Leary, A., Scott, C. L., Serra, V., Lord, C. J., Bowtell, D., et al. (2020). ESMO recommendations on predictive biomarker testing for homologous recombination deficiency and PARP inhibitor benefit in ovarian cancer. *Ann. Oncol.* 31 (12), 1606–1622. doi:10.1016/j.annonc.2020.08.2102

Mirza, M. R., Monk, B. J., Herrstedt, J., Oza, A. M., Mahner, S., Redondo, A., et al. (2016). Niraparib maintenance therapy in platinum-sensitive, recurrent ovarian cancer. *N. Engl. J. Med.* 375 (22), 2154–2164. doi:10.1056/NEJMoa1611310

Moore, K. N., Secord, A. A., Geller, M. A., Miller, D. S., Cloven, N., Fleming, G. F., et al. (2019). Niraparib monotherapy for late-line treatment of ovarian cancer (QUADRA): A multicentre, open-label, single-arm, phase 2 trial. *Lancet. Oncol.* 20 (5), 636–648. doi:10.1016/S1470-2045(19)30029-4

Nguyen, L., Martens, W. M. J., Van Hoeck, A., and Cuppen, E. (2020). Pan-cancer landscape of homologous recombination deficiency. *Nat. Commun.* 11 (1), 5584. doi:10.1038/s41467-020-19406-4

O'Kane, G. M., Connor, A. A., and Gallinger, S. (2017). Characterization, detection, and treatment approaches for homologous recombination deficiency in cancer. *Trends Mol. Med.* 23 (12), 1121–1137. doi:10.1016/j.molmed.2017.10.007

Pellegrino, B., Musolino, A., Llop-Guevara, A., Serra, V., De Silva, P., Hlavata, Z., et al. (2020). Homologous recombination repair deficiency and the immune response in breast cancer: A literature review. *Transl. Oncol.* 13 (2), 410–422. doi:10.1016/j.tranon.2019.10.010

Popova, T., Manié, E., Rieunier, G., Caux-Moncoutier, V., Tirapo, C., Dubois, T., et al. (2012). Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res.* 72 (21), 5454–5462. doi:10.1158/0008-5472.CAN-12-1470

Pujade-Lauraine, E., Ledermann, J. A., Selle, F., Gebski, V., Penson, R. T., Oza, A. M., et al. (2017). Olaparib tablets as maintenance therapy in patients with platinum-sensitive, relapsed ovarian cancer and a BRCA1/2 mutation (SOLO2/ENGOT-Ov21): A double-blind, randomised, placebo-controlled, phase 3 trial. *Lancet. Oncol.* 18 (9), 1274–1284. doi:10.1016/S1470-2045(17)30469-2

Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn.* 1 (1), 81–106. doi:10.1023/A:1022643204877

Radhakrishnan, S. K., Jette, N., and Lees-Miller, S. P. (2014). Non-homologous end joining: Emerging themes and unanswered questions. *DNA Repair* 17, 2–8. doi:10.1016/j.dnarep.2014.01.009

Ray-Coquard, I., Pautier, P., Pignata, S., Pérol, D., González-Martín, A., Berger, R., et al. (2019). Olaparib plus bevacizumab as first-line maintenance in ovarian cancer. *N. Engl. J. Med.* 381 (25), 2416–2428. doi:10.1056/NEJMoa1911361

Shen, R., and Seshan, V. E. (2016). Facets: Allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* 44 (16), e131. doi:10.1093/nar/gkw520

Sherill-Rofe, D., Rahat, D., Findlay, S., Mellul, A., Guberman, I., Braun, M., et al. (2019). Mapping global and local coevolution across 600 species to identify novel homologous recombination repair genes. *Genome Res.* 29 (3), 439–448. doi:10.1101/gr.241414.118

Swisher, E. M., Gonzalez, R. M., Taniguchi, T., Garcia, R. L., Walsh, T., Goff, B. A., et al. (2009). Methylation and protein expression of DNA repair genes: Association with chemotherapy exposure and survival in sporadic ovarian and peritoneal carcinomas. *Mol. Cancer* 8 (1), 48. doi:10.1186/1476-4598-8-48

Swisher, E. M., Lin, K. K., Oza, A. M., Scott, C. L., Giordano, H., Sun, J., et al. (2017). Rucaparib in relapsed, platinum-sensitive high-grade ovarian carcinoma

(ARIEL2 Part 1): An international, multicentre, open-label, phase 2 trial. *Lancet. Oncol.* 18 (1), 75–87. doi:10.1016/S1470-2045(16)30559-9

Sztupinszki, Z., Diossy, M., Krzystanek, M., Reiniger, L., Csabai, I., Favero, F., et al. (2018). Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. *npj Breast Cancer* 4 (1), 16. doi:10.1038/s41523-018-0066-6

Talevich, E., Shain, A. H., Botton, T., and Bastian, B. C. (2016). CNVkit: Genome-Wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput. Biol.* 12 (4), e1004873. doi:10.1371/journal.pcbi.1004873

Telli, M. L., Timms, K. M., Reid, J., Hennessy, B., Mills, G. B., Jensen, K. C., et al. (2016). Homologous recombination deficiency (HRD) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple-negative breast cancer. *Clin. Cancer Res.* 22 (15), 3764–3773. doi:10.1158/1078-0432.CCR-15-2477

Ungerboeck, G. (2006). Channel coding with multilevel/phase signals. *IEEE Trans. Inf. Theory* 28 (1), 55–67. doi:10.1109/TIT.1982.1056454

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., et al. (2013). From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* 43 (1), 11. doi:10.1002/0471250953.bi1110s43

Varma, S., and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinforma.* 7 (1), 91. doi:10.1186/1471-2105-7-91

Viola, P., Platt, J., and Zhang, C. (2007). *Multiple instance boosting for object detection*, 1417–1426.

Watkins, J. A., Irshad, S., Grigoriadis, A., and Tutt, A. N. J. (2014). Genomic scars as biomarkers of homologous recombination deficiency and drug response in breast and ovarian cancers. *Breast Cancer Res.* 16 (3), 211. doi:10.1186/bcr3670

Zhang, Q., and Goldman, S. A. (2001). EM-DD: An improved multiple-instance learning technique. *Nips* 01, 1073–1080.