Gene signature for predicting homologous recombination deficiency in triple-negative breast cancer

Jia-Wern Pan[1], Pei-Sze Ng[1], Muhammad Mamduh Ahmad Zabidi[1,2], Putri Nur Fatin[1], Jie-Ying Teo[1], Siti Norhidayu Hasan[1], Cheng-Har Yip[3], Pathmanathan Rajadurai[3,4], Lai-Meng Looi[5], Aishah M. Taib[6], Oscar M. Rueda[7], Carlos Caldas[7,8], Suet-Feung Chin[7], Joanna Lim[1], Soo-Hwang Teo[1,9]

[1]Cancer Research Malaysia, No. 1, Jalan SS12/1A, 47500 Subang Jaya, Malaysia

[2]Roche (Malaysia), The Pinnacle, Bandar Sunway, 47500 Subang Jaya, Malaysia

[3]Subang Jaya Medical Centre, No. 1, Jalan SS12/1A, 47500 Subang Jaya, Malaysia

[4]Jeffrey Cheah School of Medicine & Health Sciences, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway, 47500 Subang Jaya, Malaysia

[5]Department of Pathology, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia

[6]Department of Surgery, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia

[7]Cancer Research UK, Cambridge Institute & Department of Oncology, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK

[8]NIHR Cambridge Biomedical Research Centre and Cambridge Experimental Cancer Medicine Centre, Cambridge University Hospital NHS Foundation Trust, Cambridge, UK

[9]University Malaya Cancer Research Institute, Faculty of Medicine, University Malaya, 50603 Kuala Lumpur, Malaysia

Abstract

Triple-negative breast cancers (TNBCs) are a subset of breast cancers that has remained difficult to treat. One of the more efficient treatments for TNBCs is by targeting the homologous recombination (HR) pathway, for example using PARP inhibitors, that has benefited TNBC patients carrying *BRCA1/2* alterations. We developed a method to identify TNBC patients, regardless of their *BRCA1/2* status, who may benefit from therapies targeting the HR pathway. Using an RNA-seq gene expression dataset of 100 genes (HRD100) from the Malaysian Breast Cancer (MyBrCa) cohort, we developed a nearest centroid classifier for homologous recombination deficiency (HRD) in TNBCs. The HRD100 classifier identified samples with strong HRD mutational signature at an AUROC of 0.892 in the MyBrCa training dataset, as well as 0.783 and 0.713 in MyBrCa and TCGA validation datasets, respectively. Analysis of the 100 genes in the HRD100 classifier using the NanoString nCounter platform showed a concordance rate of 98% (CI: 95-100%) with RNA-seq gene expression analyses, and a concordance rate of 87% (CI: 73-100%) between FFPE and fresh frozen tissue. Taken together, gene expression using these 100 selected genes may identify triple-negative breast cancer patients with homologous recombination deficiency who may benefit from treatment with PARP inhibitors or platinum chemotherapy.

Introduction

Triple negative breast cancer (TNBC) continues to be an area of unmet clinical need, as this aggressive subtype has fewer treatment options and higher mortality rates compared to other subtypes of breast cancer[1–3]. As a "catch-all" diagnoses for all breast tumours that test negative for both hormone receptors and HER2, TNBC is highly heterogeneous, and can be divided into several subtypes of its own[4,5], each with potentially different treatment options. In this context, one potential biomarker that may be useful for subcategorizing TNBC patients for treatment is homologous recombination deficiency (HRD).

Cells with homologous recombination deficiency (HRD) have defects in their homologous recombination pathway leading to a diminished ability to repair DNA damage. In breast cancer, HRD is an important biomarker for therapies that utilize DNA-damaging agents such as platinum chemotherapy and PARP inhibitors. Platinum chemotherapies such as cisplatin, oxaliplatin, and carboplatin induce crosslinking of DNA that inhibits DNA repair and synthesis[6]. PARP inhibitors, on the other hand, are a new class of drugs that inhibit the action of Poly (ADP-ribose) polymerase (PARP) proteins, leading to double strand DNA breaks during cellular replication[7]. In tumours with HRD, these DNA-damaging agents cause an accumulation of mutations, leading to synthetic lethality and eventually cell death[8]. Tumours with genomic *BRCA* mutations have defective homologous recombination repair pathways[9], and PARP inhibitors have been approved for TNBC for germline carriers of deleterious variants in *BRCA1/2* (gBRCAm carriers)[10,11].

Besides germline *BRCA* mutations, other molecular alterations in tumours may also lead to similar defects in the HRD pathway. This "BRCAness" feature has been proposed to broaden the patient population to PARP inhibitors[12], and various molecular aspects of "BRCAness" have been characterized[13]. Some breast tumours with the "BRCAness" feature may arise when expressions of *BRCA1* or *BRCA2* are repressed by hypermethylation or somatic mutation[14], or when the homologous recombination pathway is abrogated through mutations in other genes in the pathway (e.g. *PALB2* and *ATM*)[15,16], and there are ongoing clinical studies that seek to expand the utility of PARP inhibitors in this context[17]. In addition, transcriptional signatures and genomic mutational signatures have been generatedto identify tumours with HRD, with some association with PARP inhibitor sensitivity[18,19]. Cinical studies to examine the utility of these genomic signatures as predictive biomarkers have also been initiated[20]. In other cancer settings such as recurrent or high grade serous ovarian carcinoma, HRD assays such as the myChoice HRD assay and the Foundation Medicine T5 NGS assay have demonstrated some utility in guiding treatment with PARP inhibitors, and have received FDA approval as companion diagnostics[21–23].

We have recently characterized the genomic and transcriptomic profiles of 560 fresh frozen breast tumours from Malaysian patients of Chinese, Malay and Indian descent (the MyBrCa cohort), including 94 TNBC samples[24]. In order to study transcriptomic biomarkers for HRD in Asian TNBC, we used the transcriptomic data from these TNBC samples and determined their associations with HRD status. We identified a set of largely novel genes that were associated with HRD in our cohort. Using this gene set, we further developed a gene-expression-based classification tool ("HRD100") to classify patient tumour samples as having high or low HRD, and validated the classifier in validation datasets from the MyBrCa and TCGA cohorts. We also validated the classifier on an alternative NanoString platform, as well as using FFPE instead of fresh frozen tumour samples. This classifier may have clinical utility as a non-gBRCAm biomarker to select for patients with high HRD who may benefit from treatment with platinum chemotherapy or PARP inhibitors.

Results

Study population

Our discovery and training dataset consisted of 94 TNBC samples from the MyBrCa cohort, for which whole-exome sequencing and RNA-seq data were derived from fresh frozen primary tumour tissue and matched blood samples (Pan et al. 2020). Of these, eight samples had pathogenic germline *BRCA1* mutations, two samples had pathogenic germline *BRCA2* mutations, and two samples had pathogenic germline *PALB2* mutations[25]. Our validation cohorts comprised of (a) an additional 35 TNBC samples from the same MyBrCa cohort that were recruited and sequenced separately; and (b) 87 TNBC tumours in The Cancer Genome Atlas (TCGA) breast cancer cohort[26] (TCGA 2012).

Hierarchical clustering of TNBC mutational signatures

First, we examined the prevalent mutational drivers in the Asian TNBC setting by conducting an unsupervised clustering analysis of the mutational signatures of the MyBrCa TNBC samples in the discovery group. Mutational signatures are specific sets of mutations, both in terms of mutation type as well as frequency, which are characteristically generated by specific mutagenic processes, which were discovered in part by statistical deconvolution of mutational data from large-scale whole-genome sequencing of tumour samples[27,28]. Using a collection of these signatures in the COSMIC database (https://cancer.sanger.ac.uk/signatures/), we deconvoluted the mutational profile of each individual tumour sample to quantitatively ascribe to each sample the mutagenic processes that were important in their development.

Hierarchical clustering of COSMIC mutational signatures for the MyBrCa tumour genomics cohort[24] revealed two large clusters of TNBC tumours with high proportions of the COSMIC mutational signature SBS3, as well as high genomic scar signature scores (HRD score), both of which have been previously associated with defective homologous recombination repair (Figure 1). These clusters included all of the tumours arising in known carriers of germline *BRCA1*/*BRCA2*/*PALB2* mutations. The results suggest that HRD is a significant driver of tumour mutations in Asian TNBC, even in the absence of germline *BRCA* mutations.

Identification of the HRD100 gene set

Given that the clusters of samples that encompassed samples with germline *BRCA* mutations are defined primarily by their SBS3 mutational signature score, we next classified the 94 MyBrCa TNBC tumours according to the proportion of somatic mutations in each tumour that correspond to the COSMIC mutational signature SBS3. Using a cutoff of 0.2 based on the distribution of SBS3 scores

in the cohort (Figure 2A), tumours with scores above the cutoff were classified as "HRD High", while tumours below the cutoff were classified as "HRD Low". To confirm that these groups of samples reflected true differences in homologous recombination repair, we also compared other commonly accepted features of HRD such as genomic loss of heterozygosity (LOH), telomeric allelic imbalance (TAI) and large-scale state transition (LST), as well as an overall sum of these three scores (HRDsum)[18], between the two groups, and found a significant increase in all of these features for the "HRD High" group ($p=7.06e^{-12}$ for HRDsum, Supp. Fig 1). Then, using differential gene expression analysis of RNA-sequencing data from the TNBC tumour samples, we identified a set of 100 genes that are differentially expressed between the two groups by manually curating the top 50 upregulated and top 50 downregulated genes based on Benjamini-Hochberg adjusted p-values, with a minimum $\log_2$ fold change of 0.5 in either direction (Figure 2B, Supp. Table 1). Gene set variation analysis (GSVA) of these 100 genes confirmed that the expression of these genes, scored for each individual sample as the difference between its GSVA scores for the upregulated and downregulated gene sets, was significantly different between the HRD High and HRD Low samples ($p=4.36e^{-7}$, Figure 2C).

Over-representation analysis of the selected 100 differentially expressed genes using KEGG and Reactome pathway-based gene sets revealed an enrichment only in the peroxisomal lipid metabolism pathway (FDR q-value < 0.01, Supp. Table 2), while an analysis of over-represented gene ontology (GO) terms indicated the involvement of metabolic and cell differentiation processes (Supp. Table 3). Similarly, gene set enrichment analysis (GSEA) of MSigDB Hallmark and KEGG pathways comparing TNBC tumour samples with high SBS3 scores to low SBS3 scores revealed an upregulation of the cell cycle pathways and downregulation of metabolic pathways in the samples with high SBS3 scores (Supp. Table 4-5).

Classification of tumour samples according to HRD status using the gene set

Using the set of 100 genes identified above, we trained a machine learning classifier to classify the HRD status of any given tumour sample, using the nearest centroid methodology described in Tibshirani et al. (2002)[29]. We first determined the standardized centroids for each of the two classes (HRD High and HRD Low) in a training dataset consisting of the 94 MyBrCa TNBC samples from the discovery group. Then, to determine each sample's HRD status, we used the median-centered, "Trimmed Mean of M-values" (TMM)-normalized transcript count data to find each sample's standardized squared distance to each centroid. This distance is then used to assign a probability to each sample as being HRD High or HRD Low using Gaussian Discriminant Analysis. We considered a sample to be unambiguously classified when the probability of the sample being in one class exceeds the other by 0.5, with ambiguous samples classified as "NA".

Using this method, the classifier classified the 45 out of the 55 Asian TNBC in the training samples with high SBS3 scores correctly as "HRD High", and classified 26 out of 39 of the samples with low SBS3 scores correctly as "HRD Low", for a combined accuracy of 75.5%. The classifier also had an AUROC of 0.892 ($p=9.03e^{-11}$) when comparing the assigned probability of each sample being "HRD High" to their actual HRD classification based on SBS3 scores (Figure 3A). The classifier was also able to categorize 10 out of the 12 samples in the training cohort with known germline *BRCA1/BRCA2/PALB2* mutations as HRD High (Figure 3B). Finally, the samples classified as "HRD

High" also had significantly higher LOH, TAI, LST, and HRDsum scores compared to samples classified as "HRD Low" (Supp. Fig. 2).

Validation of the HRD100 classifier

Next, we validated the HRD100 classifier using an additional 35 TNBC samples from the MyBrCa cohort that were collected and sequenced at a later point. Using RNA-seq gene expression data for the 100 genes, the classifier was able to select for TNBC samples with high SBS3 scores with an AUROC of 0.783 (p=0.00357), Figure 3C), correctly classifying 28 out of 35 samples as HRD High or Low according to the 0.2 cutoff (accuracy of 80%). There was also a significant difference in SBS3 scores between the MyBrCa TNBC samples in the validation cohort that were classified as HRD High vs. HRD Low (p=0.008, Figure 3D).

The classifier was further validated using data from TNBC tumours in The Cancer Genome Atlas (TCGA) dataset (TCGA 2012). Using RNA-seq gene expression data for the 100 genes, the HRD100 classifier was able to select for TCGA TNBC samples with high SBS3 scores with an AUROC of 0.713 ($p=6.17e^{-4}$, Figure 4A), correctly classifying 58 out of 87 samples as HRD High or Low according to the 0.2 cutoff (accuracy of 66.7%). There was also a significant difference in SBS3 scores between TCGA TNBC samples that were classified as HRD High vs. HRD Low ($p=1.15e^{-4}$, Figure 4B). Additionally, the HRD100 classifier was able to classify 4 out of 5 of the TCGA TNBC samples known to carry gBRCA mutations as being HRD High (Figure 4C). The lower performance of the classifier with TCGA data is perhaps to be expected, given the known differences between the MyBrCa and TCGA breast cancer cohorts[24].

Additionally, we also explored the use of the HRD classifier in a non-TNBC setting. To do so, we used gene expression and mutational signature data from 458 breast tumour samples that were estrogen, progesterone, or Her2 receptor-positive from the MyBrCa cohort[24]. Of these 458 samples, 381 had SBS3 scores less than 0.2, and 77 had SBS3 scores more than 0.2. Using gene expression data for these samples that was normalized to the median of the TNBC training cohort, the HRD100 classifier was able to select for samples with high SBS3 scores at an AUROC of 0.658 ($p=4.47e^{-5}$), correctly classifying 343 out of the 381 samples with low SBS3 scores as "HRD Low", but only classifying 13 of the 77 samples with high SBS3 scores as "HRD High" (Supp. Fig. 3).

Validation of the HRD classifier using a NanoString platform

Next, we evaluated the robustness of the HRD100 classifier with regards to different methods of measuring gene expression, as well as to the use of formalin-fixed paraffin-embedded (FFPE) tissue rather than fresh frozen tissue. We did this because both RNA-seq data as well as fresh frozen tumour tissue are expensive and difficult to obtain as part of routine clinical practice, especially in low-resource settings.

To evaluate the performance of the HRD100 classifier with a different gene expression measurement method, we used the NanoString nCounter platform[30], which uses direct digital detection of mRNA molecules to generate gene-level transcript counts. Accordingly, we designed a custom NanoString

nCounter CodeSet corresponding to the HRD100 gene set, and used the custom CodeSet to generate gene expression data from 61 fresh frozen tissue samples from the same cohort of TNBC patients that were included in the training data. The NanoString data were then inputted into the HRD100 classifier, and the results were compared to the HRD100 results from the original RNA-seq data. We found that 60 out of 61 of the NanoString fresh frozen samples corresponded with the RNA-seq results (98% concordance) (Table 1).

Finally, we repeated the NanoString experiments using 23 FFPE breast tumour tissue samples that were collected during the same surgery as the fresh frozen material, from the same cohort of TNBC patients. Out of the 23 FFPE samples, 20 of them had the same HRD100 classification results as their fresh frozen counterparts, for a concordance rate of 87% (Table 2).

Discussion

Here, we describe the development and validation of a gene expression classifier for homologous recombination deficiency in Asian TNBC. Using a nearest-centroid algorithm with transcriptomic data from the MyBrCa cohort, the HRD100 classifier was able to discriminate between samples with high versus low mutational signature scores suggestive of HR deficiency with good accuracy and AUROC in both a training and validation dataset. The HRD100 classifier had a modest but statistically significant ability to discriminate between samples with high and low HRD in a TCGA TNBC validation dataset from an entirely different population, as well as in non-TNBC samples from the MyBrCa cohort. Importantly, we also found a very high concordance rate in the classification results when using an alternative measure of gene expression (NanoString), as well as when using FFPE material instead of fresh frozen tissue, suggesting that the HRD100 classifier could be robust for real-world clinical use situations.

In the TNBC setting, we found clusters of samples with high HRD-associated genomic scores and mutational signatures. These results suggest that HRD is a significant driver of tumour mutations in Asian TNBC, even in the absence of germline *BRCA* mutations. This in turn suggests that there may be a significant number of Asian TNBC patients lacking germline *BRCA* mutations who may still benefit from therapies that target the HRD pathway, such as platinum chemotherapy or PARP inhibitors. Thus, tools to detect HRD in tumour samples may have clinical utility as a non-gBRCAm biomarker to select for Asian breast cancer patients with high HRD who may benefit from treatment with platinum chemotherapy or PARP inhibitors.

The HRD classification tool described in this paper relies on predicting the amount of mutational signature SBS3 in each tumour sample using machine learning. The COSMIC mutational signature SBS3 is a well-studied cancer mutational signature that has been validated in orthogonal techniques[27,31], and its HRD-driven etiology has been experimentally confirmed[32]. The use of HRD-associated mutational signatures to predict “BRCAness” and thus response to PARP inhibitors and platinum chemotherapy has received significant attention in recent years, with tools such as HRDetect[19] and CHORD[33] demonstrating that HRD-associated mutational signatures can be used to predict BRCA1/BRCA2-deficiency in breast tumours as well as survival of breast cancer patients[34]. The clinical utility of such tools in a breast cancer setting is being tested in ongoing clinical trials; however, one significant drawback is that these methods usually require whole-genome sequencing of

tumour samples, which may be prohibitively expensive and time-consuming, particularly in low-resource settings. By using gene expression signatures, our method for predicting HRD in tumour samples may offer a useful alternative.

Other gene expression signatures related to HRD have also been previously described - Peng and colleagues (2014)[35] described a set of 230 genes associated with homologous recombination DNA repair in a microarray analysis of a nonmalignant human mammary epithelial cell line that predicts clinical outcome in cancer patients, while a different 77-gene panel of "*BRCA1*ness" was significantly associated with response to PARP inhibitor treatment in the I-SPY 2 breast cancer clinical trial[36,37]. Interestingly, there appears to be little overlap between the different gene panels – only two out of the 100 genes in our study are included in Peng's 230-gene set, and only 8 of the 100 genes in our study are included in the 77-gene "BRCA1ness" panel, with the *TTK* gene being the only gene common across all three sets. The lack of overlap may reflect differences in methodology or study population, and further study to reconcile these differences may be warranted.

Taken together, we believe that the HRD100 classifier, implemented as a NanoString-based test, may have clinical utility as a non-*BRCA*m biomarker to select for patients with high HRD who may benefit from treatment with PARP inhibitors. However, further development of the HRD100 classifier will require the initiation of clinical trials in order to determine if HRD100 can correctly identify patients who are sensitive PARP inhibitor therapy in real-world clinical settings.

Methods

Data description

Genomic sequencing data for this project was taken primarily from TNBC samples included in the MyBrCa cohort tumour sequencing project. In brief, this included whole-exome sequencing (WES) and RNA-sequencing data collected from breast tumours of female patients from two hospitals – Subang Jaya Medical Centre in Subang Jaya, Malaysia, and Universiti Malaya Medical Centre in Kuala Lumpur, Malaysia, and analysed together with available clinical data. The cohort data and sequencing methods are described in full in Pan et al. (2020) and associated papers[24,25,38]. Validation data from TCGA TNBC samples were downloaded from the NIH Genomics Data Portal.

Transcriptomic data processing

Raw RNA-Seq reads were mapped to the hs37d5 reference human genome, and gene-level read counts were quantified using featureCounts (v. 1.2.31) with the Homo sapiens GRCh37.87 human transcriptome genome annotation model.

Mutational analyses

To call SNVs, we used positions called by Mutect2 with following filters: minimum 10 reads in tumour and 5 reads in normal samples, OxoG metric less than 0.8, variant allele frequency (VAF) 0.075 or more, p-value for Fisher's exact test on the strandedness of the reads 0.05 or more, and S_{AF} more than 0.75. For positions that are present in 5 samples or more, we removed two positions that were not in COSMIC and in single tandem repeats. We also removed variants that have VAF at least 0.01 in gnomAD, and considered only variants that are supported by at least 4 alternate reads, with at least 2 reads per strand. For indels, we also required the positions to be called by Strelka2. Variants were annotated using Oncotator version 1.9.9.0. To determine the prevalence of previously-reported breast cancer single base-pair substitution (SBS) mutational signatures from COSMIC matrices (Signatures 1, 2, 13, 3, 8, 6, 15, 20, 26, 5, 17, 18 and 30), we used deconstructSigs[39] (v.1.8.0), restricted to samples with at least 15 SNVs.

Statistical analyses

Hierarchical clustering of mutational signatures was done using the "hclust" function implemented in the base stats package (v. 3.6.2) in R. All box and whiskers plots in the figures are constructed with boxes indicating 25th percentile, median and 75th percentile, and whiskers showing the maximum and minimum values within 1.5 times the inter-quartile range from the edge of the box, with outliers not shown.

Differential expression analyses

Gene-level count matrices were normalised using the "Trimmed Mean of M-values" method implemented in the edgeR (v. 3.20.9) R package. The count matrices were then transformed into logCPM using the voom function from the limma package in R. Differentially expressed genes were determined by empirical Bayes moderation of the standard errors towards a common value from a linear model fit of the transformed count matrices as implemented in the limma package, with the

threshold for differential expression set as false discovery rate (FDR) < 0.001 and absolute log fold change >0.2.

Pathway analysis

Over-representation analysis using KEGG and Reactome pathway-based sets as well as gene-ontology (GO) based sets was conducted using ConsensusPathDB (http://cpdb.molgen.mpg.de, accessed 21 April 2022) using the human database and ENSEMBL identifiers. For GO-based sets, the search was restricted to gene ontology level 2 and level 3 categories only.

Pathway analysis was conducted using gene set enrichment analysis (GSEA), as implemented in the Broad Institute GSEA Java executable (v 4.2.3), using the MSigDB Hallmark gene sets, as well as the KEGG gene sets, as implemented in the GSEA program using default options.

Determination of germline *BRCA* mutation status

Carriers of deleterious germline mutations in *BRCA1* and *BRCA2* in the MyBrCa cohort were identified from targeted sequencing conducted as part of the BRIDGES study[40]. Each carrier was independently confirmed with Sanger sequencing. Comparable data for the TCGA TNBC cohort was obtained from Huang et al.[41].

Nearest-centroid classification

Nearest-centroid classification was conducted as implemented in the "pamr" R package (v. 1.56.1) using default options. The training dataset consists of RNA-seq gene expression data quantified as TMM and $\log_2$ normalized counts per million (CPM), along with the HRD classification of each sample. RNA-seq read counts from the validation datasets were also normalized in the same way.

RNA extraction

RNA from tumour samples was extracted using the QIAGEN miRNeasy Mini Kit with a QIAcube, according to standard protocol. Total RNA was quantitated using a Nanodrop 2000 Spectrophotometer and RNA integrity was measured using an Agilent 2100 Bioanalyzer.

NanoString validation

Using the NanoString nCounter platform, we developed a custom CodeSet in order to measure the expression of 100 genes – 97 genes from our differential expression analysis and 3 housekeeping genes used for data normalization. Expression for this gene set was measured on an nCounter MAX Analysis System, and the raw data was processed and normalized using the NanoString's proprietary nSolver (v. 4.0) software before being exported as a text file for processing by the machine learning classifier.

Acknowledgements

Cancer Research Malaysia receives charitable funding from the Scientex Foundation, Estée Lauder Companies, Vistage Malaysia, Yayasan Petronas, and Yayasan Sime Darby which contributed to the funding of this study. OMR, CC, and SFC also receive funding from Cancer Research UK. All genomics work was undertaken by the Genomics Core Facility CRUK Cambridge Institute.

Author Contributions

JWP led the data analysis, supervised experiments, and wrote the manuscript. PSN contributed to data analysis, experimental design, sample collection, and carried out experiments. MMAZ and PNF contributed to data analysis and generation of figures. JYT, SNH, CHY, PR, LML, and AMT contributed to sample collection and processing and data collection, while OMR and SFC generated and processed sequencing data. PR and LML provided histopathology expertise, and together with CHY and AMT collected clinical data. OMR, CC, and SFC also interpreted results and helped to draft the manuscript. CC and SHT contributed to obtaining funding for the project. SHT also designed experiments, drafted the manuscript, and provided overall project direction and supervision.

Data Accessibility

The data that support the findings of this study are available on the European Genome-phenome Archive under the study accession number EGAS00001004518. Access to controlled patient data will require the approval of the Data Access Committee.

References

1. Yin, L., Duan, J. J., Bian, X. W. & Yu, S. C. Triple-negative breast cancer molecular subtyping and treatment progress. *Breast Cancer Res.* **22**, 1–13 (2020).

2. Lebert, J. M., Lester, R., Powell, E., Seal, M. & McCarthy, J. Advances in the Systemic Treatment of Triple-Negative Breast Cancer. *Curr. Oncol. 2018, Vol. 25, Pages 142-150* **25**, 142–150 (2018).

3. Dawson, S. J., Provenzano, E. & Caldas, C. Triple negative breast cancers: clinical and prognostic implications. *Eur. J. Cancer* **45 Suppl 1**, 27–40 (2009).

4. Lehmann, B. D. *et al.* Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.* **121**, 2750–2767 (2011).

5. Burstein, M. D. *et al.* Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clin. Cancer Res.* **21**, 1688–1698 (2015).

6. Rabik, C. A. & Dolan, M. E. Molecular Mechanisms of Resistance and Toxicity Associated with Platinating Agents. *Cancer Treat. Rev.* **33**, 9 (2007).

7. Rouleau, M., Patel, A., Hendzel, M. J., Kaufmann, S. H. & Poirier, G. G. PARP inhibition: PARP1 and beyond. *Nat. Rev. Cancer* **10**, 293 (2010).

8. O'Connor, M. J. Targeting the DNA Damage Response in Cancer. *Mol. Cell* **60**, 547–560 (2015).

9. Powell, S. N. & Kachnic, L. A. Roles of BRCA1 and BRCA2 in homologous recombination, DNA replication fidelity and the cellular response to ionizing radiation. *Oncogene 2003 2237* **22**, 5784–5791 (2003).

10. Robson, M. *et al.* Olaparib for Metastatic Breast Cancer in Patients with a Germline *BRCA* Mutation. *N. Engl. J. Med.* **377**, 523–533 (2017).

11. Litton, J. K. *et al.* Talazoparib in Patients with Advanced Breast Cancer and a Germline *BRCA* Mutation. *N. Engl. J. Med.* **379**, 753–763 (2018).

12. Lord, C. J. & Ashworth, A. BRCAness revisited. *Nat. Rev. Cancer* **16**, 110–120 (2016).

13. Severson, T. M. *et al.* BRCA1-like signature in triple negative breast cancer: Molecular and clinical characterization reveals subgroups with therapeutic potential. *Mol. Oncol.* **9**, 1528–1538 (2015).

14. Nicolas, E., Bertucci, F., Sabatier, R. & Gonçalves, A. Targeting BRCA Deficiency in Breast Cancer: What are the Clinical Evidences and the Next Perspectives? *Cancers (Basel).* **10**, (2018).

15. Wu, S. *et al.* Molecular Mechanisms of PALB2 Function and Its Role in Breast Cancer Management. *Front. Oncol.* **10**, 301 (2020).

16. Ahlskog, J. K., Larsen, B. D., Achanta, K. & Sørensen, C. S. ATM/ATR□mediated phosphorylation of PALB2 promotes RAD51 function. *EMBO Rep.* **17**, 671 (2016).

17. Gruber, J. J., Gross, W., McMillan, A., Ford, J. M. & Telli, M. L. A phase II clinical trial of talazoparib monotherapy for PALB2 mutation-associated advanced breast cancer. *https://doi.org/10.1200/JCO.2021.39.15_suppl.TPS1109* **39**, TPS1109–TPS1109 (2021).

18. Telli, M. L. *et al.* Homologous Recombination Deficiency (HRD) Score Predicts Response to Platinum-Containing Neoadjuvant Chemotherapy in Patients with Triple Negative Breast Cancer. *Clin. Cancer Res.* **22**, 3764–3773 (2016).

19. Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).

20. Chopra, N. *et al.* Homologous recombination DNA repair deficiency and PARP inhibition activity in primary triple negative breast cancer. *Nat. Commun.* **11**, (2020).

21. Haunschild, C. E. & Tewari, K. S. Bevacizumab use in the frontline, maintenance and recurrent settings for ovarian cancer. *Future Oncol.* **16**, 225–246 (2020).

22. Ray-Coquard, I. *et al.* Olaparib plus Bevacizumab as First-Line Maintenance in Ovarian Cancer. *N. Engl. J. Med.* **381**, 2416–2428 (2019).

23. Banerjee, S. *et al.* Maintenance olaparib for patients with newly diagnosed advanced ovarian cancer and a BRCA mutation (SOLO1/GOG 3004): 5-year follow-up of a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet Oncol.* **22**, 1721–1731 (2021).

24. Pan, J. W. *et al.* The molecular landscape of Asian breast cancers reveals clinically relevant population-specific differences. *Nat. Commun.* **11**, 1–12 (2020).

25. Ng, P. S. *et al.* Characterisation of PALB2 tumours through whole-exome and whole-transcriptomic analyses. *NPJ breast cancer* **7**, (2021).

26. Network, C. G. A. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).

27. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

28. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–93 (2012).

29. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 6567–6572 (2002).

30. Kulkarni, M. M. Digital multiplexed gene expression analysis using the NanoString nCounter system. *Curr. Protoc. Mol. Biol.* **Chapter 25**, (2011).

31. Polak, P. *et al.* A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* 1–15 (2017).

32. Zámborszky, J. *et al.* Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. *Oncogene* **36**, 746–755 (2017).

33. Nguyen, L., W. M. Martens, J., Van Hoeck, A. & Cuppen, E. Pan-cancer landscape of homologous recombination deficiency. *Nat. Commun. 2020 111* **11**, 1–12 (2020).

34. Staaf, J. *et al.* Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nat. Med.* **25**, 1526–1533 (2019).

35. Peng, G. *et al.* Genome-wide transcriptome profiling of homologous recombination DNA repair. *Nat. Commun.* **5**, (2014).

36. Wolf, D. M. *et al.* DNA repair deficiency biomarkers and the 70-gene ultra-high risk signature as predictors of veliparib/carboplatin response in the I-SPY 2 breast cancer trial. *NPJ breast cancer* **3**, (2017).

37. Severson, T. M. *et al.* The BRCA1ness signature is associated significantly with response to PARP inhibitor treatment versus control in the I-SPY 2 randomized neoadjuvant setting. *Breast Cancer Res.* **19**, (2017).

38. Pan, J. *et al.* Germline <scp> *APOBEC3B* </scp> deletion increases somatic hypermutation in Asian breast cancer that is associated with Her2 subtype, <scp> *PIK3CA* </scp> mutations and immune activation. *Int. J. Cancer* ijc.33463 (2021) doi:10.1002/ijc.33463.

39. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. deconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 1–11 (2016).

40. Dorling, L. *et al.* Breast Cancer Risk Genes — Association Analysis in More than 113,000

Women. *N. Engl. J. Med.* **384**, 428–439 (2021).

41. Huang, K. lin *et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* **173**, 355-370.e14 (2018).

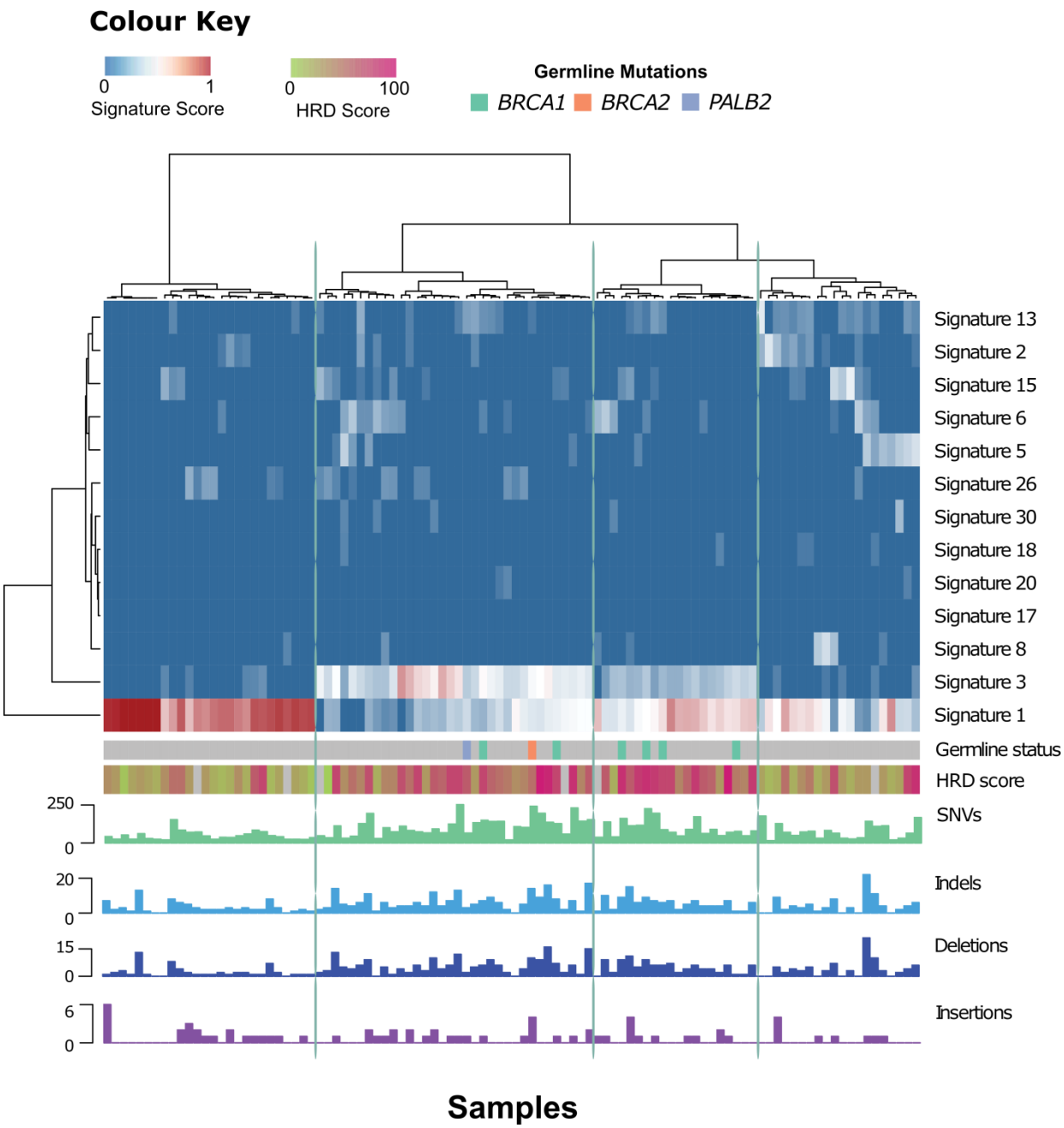

Figure 1. Cluster analysis of 94 TNBC tumours arising in Malaysian women. Four distinct clusters are observed: (left) TNBC with predominantly Signature 1; (center-left) TNBC with Signature 3 (HRD deficiency); (center-right) TNBC with Signature 1 and 3; and (right) mixed group of TNBC with predominantly signature 5, 6 and 15. Notably, whilst germline carriers of *BRCA1* and *BRCA2* predominantly are in the central clusters, the majority of these tumours arise in non carriers.

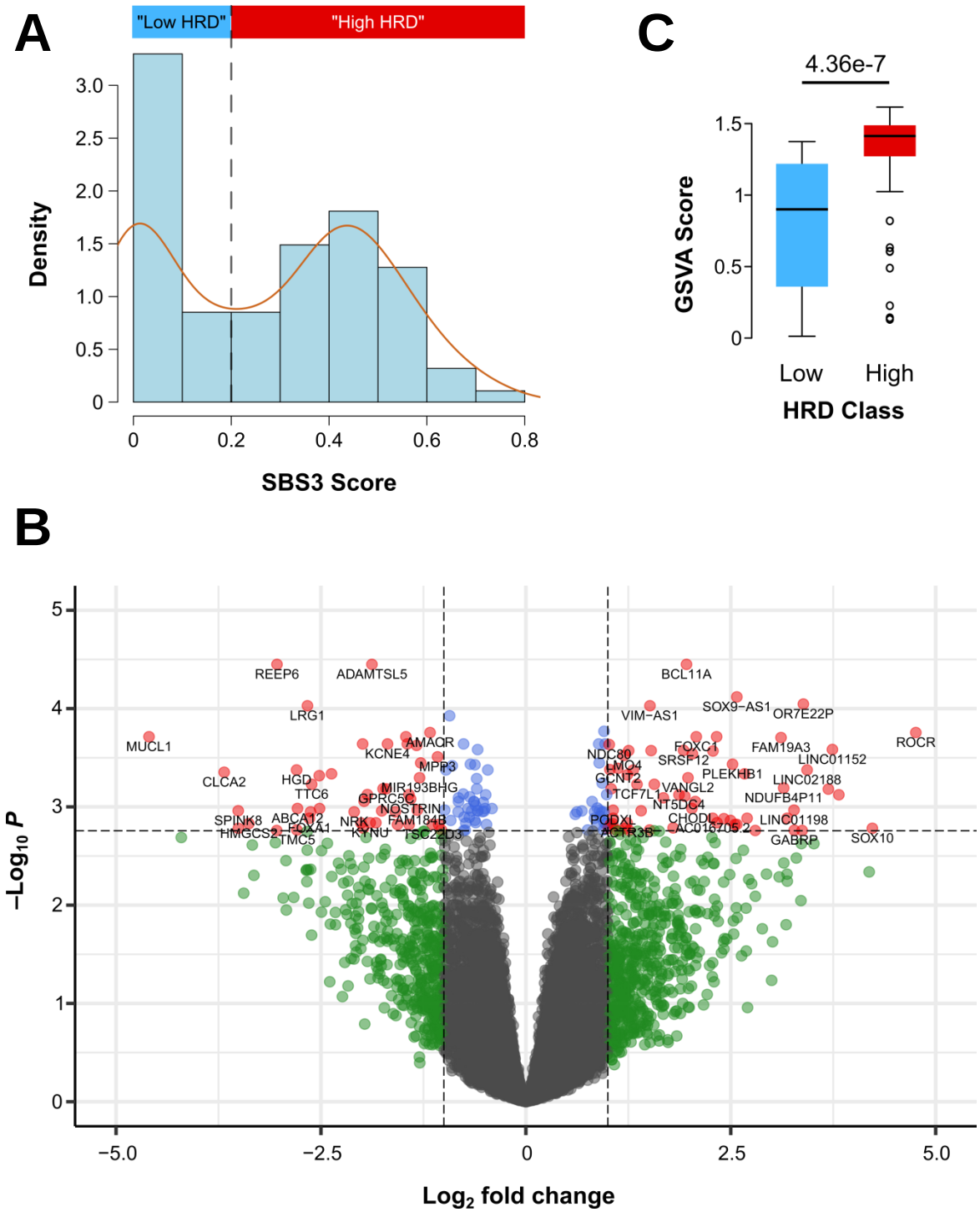

Figure 2. Differential expression analysis for HRD in 94 TNBC tumours arising in Malaysian women. (A) Density plot of the COSMIC SBS3 mutational signature in the 94 MyBrCa TNBC samples. **(B)** Volcano plot for the results of the differential expression analysis between HRD100 High vs. HRD Low samples in the 94 MyBrCa TNBC samples. **(C)** Gene Set Variation Analysis scores in HRD High and HRD Low samples for the 100 genes selected for the HRD100 classifier.

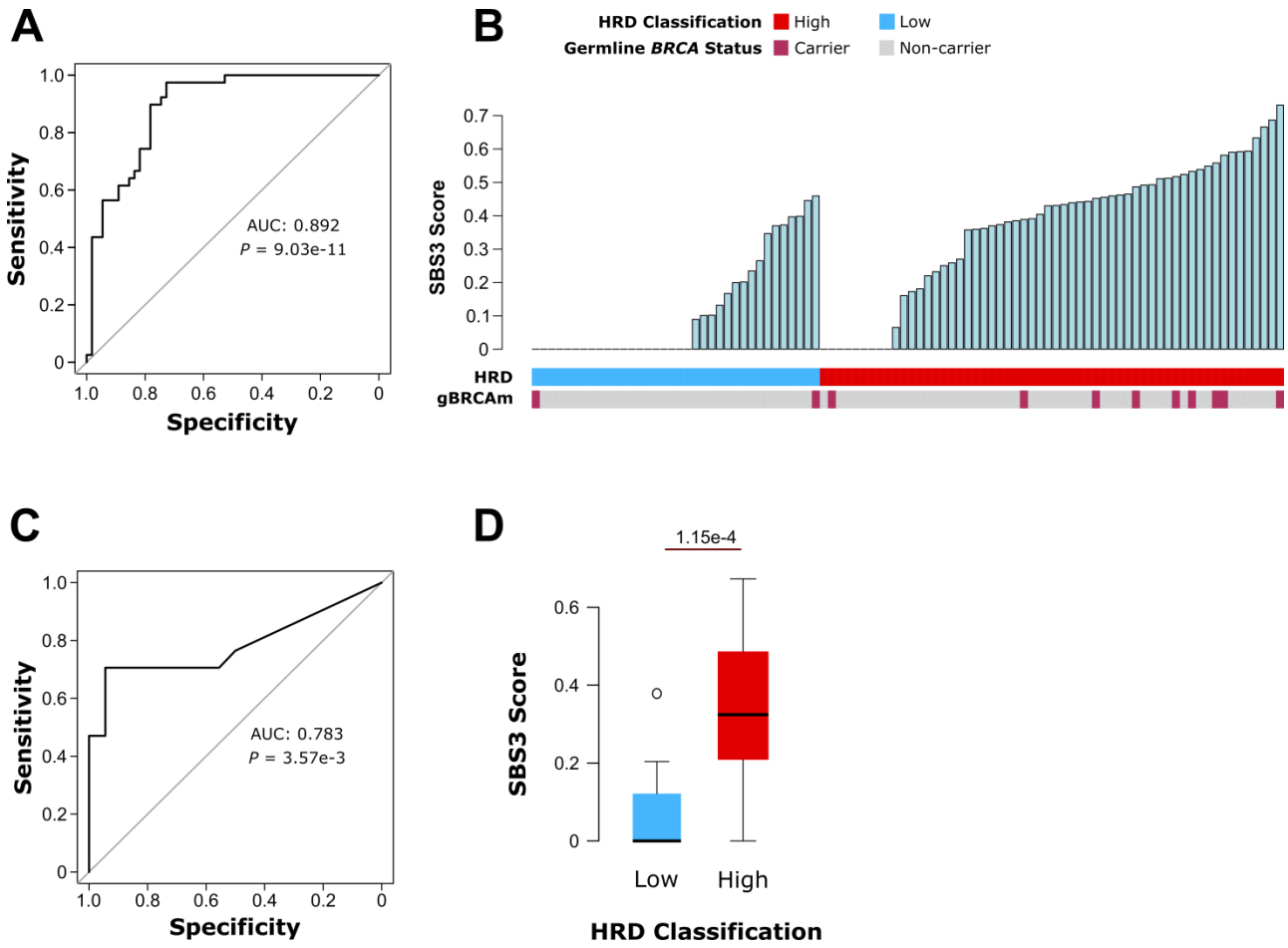

Figure 3. Performance of the HRD classifier in the training and validation datasets of the MyBrCa cohort. **(A)** ROC plot of the probability that a sample is HRD High according to the HRD100 classifier compared to actual HRD classification based on SBS3 scores in the 94 MyBrCa TNBC training dataset. **(B)** HRD100 classification of individual MyBrCa TNBC samples compared to their SBS3 scores, and samples with germline *BRCA* mutation (g*BRCA*m) are indicated in red. **(C)** ROC plot of the probability that a sample is HRD High according to the HRD100 classifier compared to actual HRD classification based on SBS3 scores in 33 MyBrCa TNBC validation samples. **(D)** SBS3 scores of the 33 validation MyBrCa TNBC samples compared to their HRD classification.

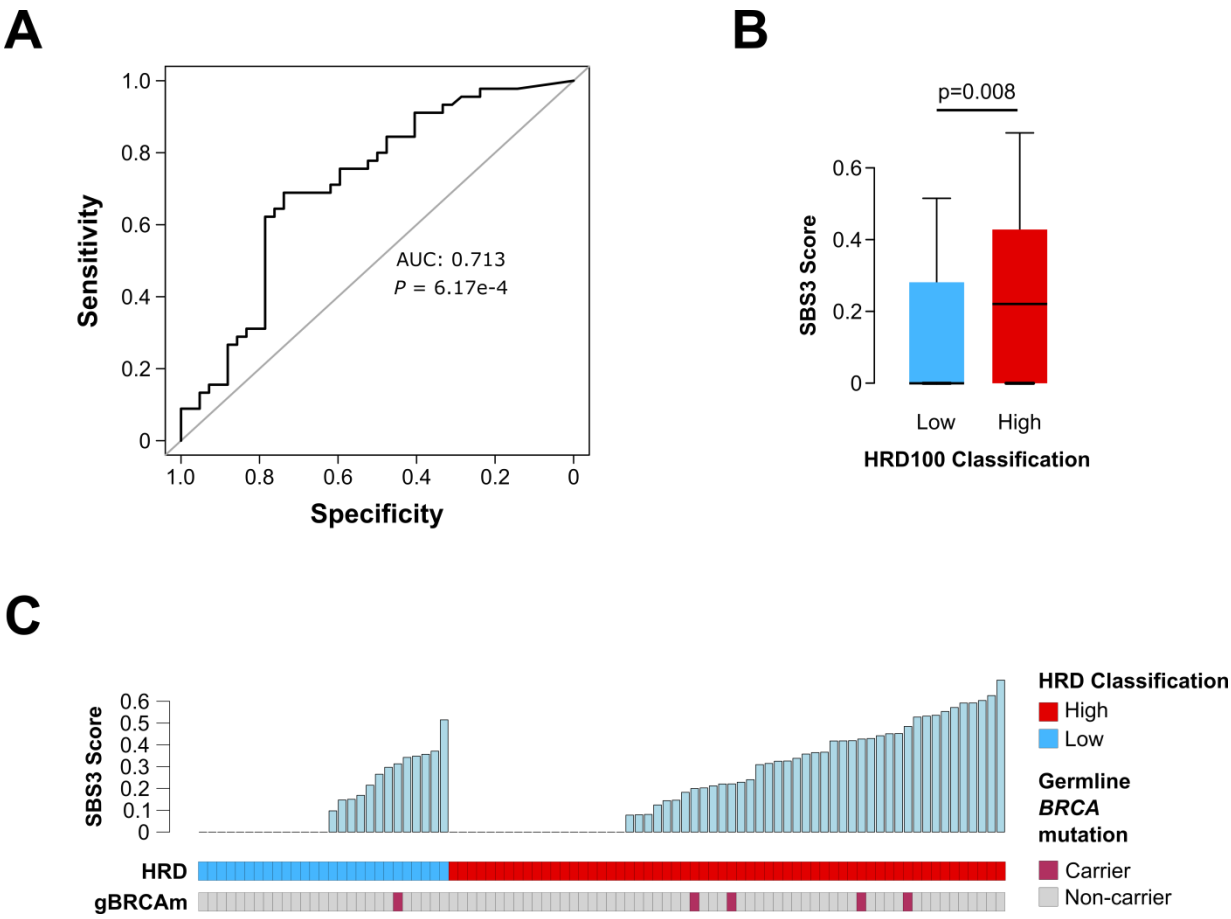

Figure 4. Validation of the HRD classifier in the TCGA cohort. (A) ROC plot of the probability that a sample is HRD High according to the HRD100 classifier compared to actual HRD classification based on SBS3 scores in 87 TCGA TNBC samples. (B) SBS3 scores of the 87 validation TCGA TNBC samples compared to their HRD classification. (C) HRD classification of individual TCGA TNBC samples compared to their SBS3 scores, and samples with germline *BRCA* mutation (g*BRCA*m) are indicated in red.

Table 1. Comparison of HRD100 classification results between RNA-seq (Illumina) and NanoString nCounter platforms using fresh frozen tissue.

		HRD100 Classification (NanoString Fresh Frozen)			
		High	Low	Inconclusive	Total
HRD100 Classification (RNA-seq)	High	37	0	0	37
	Low	0	23	1	24
	Total	37	23	1	61

Table 2. Comparison of HRD100 classification results between RNA-seq (Illumina) using fresh frozen tissue and NanoString nCounter using FFPE tissue.

		HRD100 Classification (NanoString FFPE)			
		High	Low	Inconclusive	Total
HRD100 Classification (RNA-seq)	High	13	0	1	14
	Low	1	7	1	9
	Total	14	7	2	23