UNIVERSITY OF CALIFORNIA SAN DIEGO

A machine learning approach for detecting homologous recombination deficiency in breast cancer using transcriptome data

A Thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Bioengineering

by

Mia Josephine Jeffris

Committee in charge:

Professor Ludmil B. Alexandrov, Chair
Professor Xiaohua Huang
Professor Shankar Subramaniam

2024

The Thesis of Mia Josephine Jeffris is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

To my colleagues, family and friends for their endless support and encouragement.

To the patients who generously share their data for research and participate in clinical trials, you are true heroes and our work would not be possible without your courage and selflessness.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AUC	Area under the receiver operating characteristic curve
DE	Differential expression
DNA	Deoxyribonucleic acid
DSB	Double strand break
FDR	False discovery rate
FPKM	Fragments per kilobase of transcript per million mapped reads
HR	Homologous recombination
HRD	Homologous recombination-deficient
HRD score	Homologous recombination deficiency score
HRP	Homologous recombination-proficient
LOH	Loss of heterozygosity
LST	Large-scale state transition
Mbs	Megabases
ML	Machine learning
MMEJ	Microhomology-mediated end-joining
mRNA	Messenger ribonucleic acid
NHEJ	Non-homologous end-joining
NR	Non-responder
padj	Adjusted p-value
PARP	Poly (ADP-ribose) polymerase
PARPi7	Daemen *et al.*'s 7-gene PARP inhibitor sensitivity signature (Daemen et al. 2012)
pCR	Pathological complete response

R	Responder
RNA	Ribonucleic acid
RNA-Seq	RNA-Sequencing
ROC	Receiver operating characteristic curve
SSA	Single-strand annealing
SSB	Single strand break
SVC	Support vector classifier
TAI	Telomeric allelic imbalance
TCGA-BRCA	The Cancer Genome Atlas's breast cancer cohort
TMM	Trimmed mean of M-values
TNBC	Triple-negative breast cancer
TPM	Transcripts per million

ACKNOWLEDGEMENTS

I would like to express my gratitude to Professor Ludmil B. Alexandrov for his support as my advisor and the chair of my committee. His feedback and expertise were invaluable in guiding the direction and ensuring the successful completion of this project.

I am also grateful to Ammal Abbasi for offering her support and knowledge throughout this project. Her mentorship was essential for the comprehensiveness and completion of this thesis.

I would also like to thank Dr. Erik Bergstrom for his help in identifying molecular subtype annotations for TCGA-BRCA.

VITA

2023 Bachelor of Science in Bioengineering: Bioinformatics, University of California San Diego

2024 Master of Science in Bioengineering, University of California San Diego

FIELD OF STUDY

Major Field: Bioengineering
Studies in Bioinformatics
Professor Ludmil B. Alexandrov

ABSTRACT OF THE THESIS

A machine learning approach for detecting homologous recombination deficiency in breast cancer using transcriptome data

by

Mia Josephine Jeffris

Master of Science in Bioengineering

University of California San Diego, 2024

Professor Ludmil B. Alexandrov, Chair

Breast cancer patients with deficiencies in the homologous recombination (HR) pathway are sensitive to poly (ADP-ribose) polymerase (PARP) inhibitors and platinum-based chemotherapy. Various methods have been developed to detect HR deficiency, many of which rely on genomic features indicative of impaired HR machinery. However, genomic-based

approaches cannot distinguish between current and past HR deficiency. In contrast, the transcriptome can capture the current state of homologous recombination and may be more effective than genome-based methods in reflecting the dynamic nature of the HR pathway in cancer. Recently, the clinical utility of transcriptional classifiers has been demonstrated for identifying HR-deficient (HRD) prostate cancers, but there remains a need for robust and widely-applicable transcriptome methods for detecting HR deficiency in breast cancer. In this thesis, I developed a 153-gene transcriptional signature for detecting HR deficiency in breast cancer patients, allowing identification of individuals whose cancers may be sensitive to PARP inhibitors. This signature offers an advantage over existing models due to its reduced feature set and ability to generalize across different subtypes of HRD breast cancer.

Chapter 1 Background: Breast Cancer and the Homologous Recombination DNA Repair Pathway

More than 10,000 DNA damage events occur daily in a typical human cell (Lindahl and Barnes 2000). DNA damage arises from a diverse range of processes and takes on many different forms in the genome, including abasic sites, mismatches, crosslinks, single strand breaks (SSBs), and double strand breaks (DSBs) (Carusillo and Mussolino 2020). Abasic sites are characterized by a missing nucleotide base in an otherwise continuous DNA strand, and can be generated by reactive oxygen and nitrogen species-induced oxidative damage (Van Houten et al. 2018). Mismatches occur when nucleotide bases are paired with a wrong base on the opposing DNA strand instead of adhering to adenosine-thymine or cytosine-guanine pairing. These mismatches arise as a result of erroneous DNA replication or repair (Ganai and Johansson 2016; Li 2008). Crosslinks are structural deformations characterized by the fusion of separate DNA strands, and can arise from endogenous metabolite activity as well as exogenous exposures to chemicals and radiation (Muniandy et al. 2010). SSBs are breaks that occur in a single DNA strand and can result from oxidative damage and errors in DNA replication and transcription (Caldecott 2008; Wang 2002).

DSBs are breaks across both strands of the DNA double helix. They are a particularly deleterious type of DNA lesion as they can inactivate the genes they occur in, making it so that a single DSB can trigger cell cycle arrest or cell death if occurring in an essential gene (Huang et al. 1996; Khanna and Jackson 2001). DSBs can arise endogenously as a result of interruptions in DNA replication, faulty telomere metabolism, or oxidative and mechanical damage (Khanna and Jackson 2001). Exogenous causes of DSBs include exposure to ionizing radiation and chemotherapeutic drugs (Khanna and Jackson 2001). They are formed by either the combined

effect of two independent SSBs in close proximity or by DNA replication machinery encountering an interrupting lesion, such as an SSB, on the template strand during replication (Jackson and Bartek 2009).

Cells have evolved complex DNA repair pathways over time to mitigate and reverse harmful DNA damage events and promote survival. These pathways are highly specialized for optimal repair of different types of DNA damage (Jackson and Bartek 2009). In DNA damage, there is often an intact, undamaged strand which is complementary to the damaged strand. This undamaged strand can act as a guide to re-synthesize the correct sequence at the complementary damage site. In a DSB, the breakage of both strands across the double helix means that there is no guide strand from which to rebuild the break; thus, DSB repair poses the greatest challenge for the DNA repair machinery (Khanna and Jackson 2001).

Several pathways are involved in DSB repair. One such pathway is homologous recombination, during which the DNA on either side of the DSB is resected to generate single-stranded "tails" that invade intact, double-stranded, homologous DNA molecules to serve as templates for extension by DNA polymerase (Khanna and Jackson 2001). The use of a homologous DNA double helix allows the original DNA sequence to be conserved following HR-mediated repair (Ceccaldi et al. 2016). This error-free conservation is unique to the HR pathway; thus, it is the canonical pathway for DSB repair. If the HR pathway is not functional, cells must rely on alternative error-prone pathways, such as non-homologous end-joining (NHEJ), microhomology-mediated end-joining (MMEJ), and single-strand annealing (SSA), to repair DSBs (Ceccaldi et al. 2016). Importantly, DSB repair pathway choice also depends on the cell-cycle phase at the time of repair initiation; for instance, NHEJ is dominant in the G0/G1 and G2 phases while HR is dominant in the mid-S and mid-G2 phases when a homologous chromatid

is available for use as a template (Ceccaldi et al. 2016). Furthermore, certain post-translational DNA modifications have also been shown to influence the choice of pathway for DSB repair (Ceccaldi et al. 2016).

NHEJ, MMEJ, and SSA can all lead to the formation of deleterious mutations around the DSB repair site (Ceccaldi et al. 2016). The primary mechanism of NHEJ involves ligating both ends of the DSB together (Khanna and Jackson 2001). However, if degradation has already occurred at the DSB prior to ligation, NHEJ machinery may join the DSB at microhomology sequences internal to the broken ends, introducing a small deletion at the repair site (Khanna and Jackson 2001; Ceccaldi et al. 2016; Pannunzio et al. 2014). This particular mechanism, along with the resulting microhomology-flanked deletions, has also been attributed to the MMEJ pathway (Sfeir and Symington 2015); however, under MMEJ, microhomologies are exposed through intentional end resection (Ceccaldi et al. 2016). Additionally, MMEJ has been observed to erroneously join DSBs on different chromosomes, leading to the formation of translocations and rearrangements in the genome (Ceccaldi et al. 2016). SSA induces end joining at longer homologies, specifically nucleotide repeat sites, which can result in the deletion of repeat copies as well as the intervening sequences between repeats (Ceccaldi et al. 2016).

Genomic instability is a hallmark of cancer, with defective DNA repair machinery playing a critical role in the accumulation of DNA damage (Carusillo and Mussolino 2020; Alhmoud et al. 2020). For instance, impairment of the HR pathway increases the activity of these error-prone pathways for DSB repair, ultimately resulting in the accumulation of deleterious mutations in the genome (Scully et al. 2019). Patients with a deficient HR pathway are referred to as homologous recombination-deficient (HRD), while patients with properly functioning HR pathways are generally referred to as homologous recombination-proficient (HRP). The HRD

phenotype has been observed in several cancer types, including breast, ovarian, prostate and pancreatic cancers (Stewart et al. 2022; Heeke et al. 2018; Marquard et al. 2015). HR deficiency typically results from the malfunction of genes within the HR pathway and can be detected by identifying impairments in HR genes, including germline mutations, somatic mutations, methylation changes and others (Polak et al. 2017; Stewart et al. 2022). *BRCA1* and *BRCA2* are key genes of the HR pathway, and their impairment has been consistently linked to HR deficiency in breast cancer (Stewart et al. 2022; Vollebergh et al. 2014; Heeke et al. 2018; Nguyen et al. 2020). Impairment of other HR pathway genes *ATM*, *PALB2* and *RAD51* has also been associated with HR deficiency in breast cancer, albeit less consistently than the *BRCA1/2* genes (Stewart et al. 2022; Heeke et al. 2018; Nguyen et al. 2020; den Brok et al. 2017).

Importantly, HR deficiency can also be detected by the presence of specific patterns of genomic changes (Stewart et al. 2022; Marquard et al. 2015). At least seven mutational signatures have already been associated with HR deficiency: single base substitution signatures SBS3 and SBS8 (Alexandrov et al. 2013), genome rearrangement signatures RS3 and RS5 (Nik-Zainal et al. 2016), small-scale indel signatures ID6 and ID8 (Alexandrov et al. 2020), and the CN17 copy number signature (Steele et al. 2022). Furthermore, there are three primary large-scale mutational events associated with HR deficiency in breast cancer: loss of heterozygosity (LOH), telomeric allelic imbalances (TAIs), and large-scale state transitions (LSTs) (Telli et al. 2016; Stewart et al. 2022). LOH is defined as the loss of one copy of a region of DNA. The LOH metric has been defined as the number of intermediate LOH regions longer than 15 megabases (Mbs), but shorter than a whole chromosome, in a genome sample (Abkevich et al. 2012). The TAI metric refers to the number of sub-chromosomal regions in the genome with allelic imbalance extending to the telomere but not crossing the centromere (Birkbak et al.

2012). LSTs refer to chromosomal breaks ≥ 10 Mbs between two adjacent regions in the genome, where a chromosomal break is defined as a translocation, inversion or deletion (Popova et al. 2012; Stewart et al. 2022). A high count of any of these scarring patterns has been associated with HR deficiency in breast cancer, and is thought to result from error-prone DSB repair in HRD cells (Abkevich et al. 2012; Birkbak et al. 2012; Popova et al. 2012; Stewart et al. 2022). Telli *et al.* have derived an HR deficiency score (HRD score) defined as the unweighted numeric sum of LOH, TAIs and LSTs to quantify the prevalence of HR deficiency-related genomic scarring in a sample (Telli et al. 2016). An HRD score ≥ 42 and/or *BRCA1/2* impairment were found to sufficiently detect the HRD phenotype in triple-negative breast cancer (TNBC) (Telli et al. 2016).

Tumors exhibiting the HRD genotype are vulnerable to poly (ADP-ribose) polymerase (PARP) inhibitors and platinum therapies (Tutt et al. 2018; Moore et al. 2018). Moreover, HR deficiency has been shown to confer a better response to these treatments than HR proficiency (Heeke et al. 2018; Mateo et al. 2015; Telli et al. 2016). Examples of PARP inhibitors used for breast cancer treatment include olaparib, talazoparib, rucaparib, niraparib and veliparib, and platinum-based therapies currently in use include cisplatin, carboplatin, and oxaliplatin (Cortesi et al. 2021; Zhang et al. 2022). The normal function of the PARP enzyme is to bind to DNA at SSB sites, recruit the XRCC1 DNA repair protein and then dissociate to allow access to other SSB repair factors (Curtin and Szabo 2020). PARP inhibitors prevent the dissociation of PARP from DNA, which obstructs repair machinery as well as DNA replication forks at SSB sites. (Moore et al. 2018; O'Connor 2015). This leads to the accumulation of SSBs and the generation of DSBs, which arise from the stalling and collapse of replication forks at PARP-obstructed DNA regions, in the genome (Moore et al. 2018; O'Connor 2015). The therapeutic efficacy of

PARP inhibitors in HRD tumors is attributed to the concept of synthetic lethality, wherein a defect in either the PARP or HR machinery can be tolerated by the cell, but the combination of both leads to the accumulation of deleterious mutations and subsequent cell death (Lord and Ashworth 2017; Moore et al. 2018). On the other hand, platinum therapies can introduce interstrand crosslinks to the genome, which the HR pathway also helps to repair (Wang and Lippard 2005; Deans and West 2011). For HRD tumors, this DNA damage can accumulate and eventually lead to p53-mediated apoptosis (Wang and Lippard 2005). These vulnerabilities provide an opportunity to optimize treatment for HRD cancer patients by targeting the weaknesses of HRD tumors; as such, it is important to have methods for determining the HR status of a tumor in breast and other cancers.

While *BRCA1/2* status and genomic scarring patterns are good identifiers of HR deficiency, they do not fully describe the complexity of how HR deficiency presents in the cell. These metrics only provide information about the mutational landscape of HR deficiency, which is effectively the history of all of the mutations that have ever occurred throughout the lineage of the cancer cell. This can result in the failure to accurately classify a tumor as HRD or HRP; for example, mutations which occur in the 53BP1 pathway or which restore *BRCA1* expression can lead to restoration of HR function in *BRCA1*-deficient cells and subsequent PARP inhibitor resistance (Jacobson et al. 2023; Dias et al. 2021; Noordermeer et al. 2018). Thus, one needs to define metrics for HR deficiency classification which relate to the current state of a tumor.

Determining the transcriptional profiles of known HRD and HRP tumors is one possible solution for this, as they would describe only the most recent gene activity rather than the accumulation of mutations that have occurred since the formation of the tumor. The clinical utility of transcriptional signatures for HR deficiency has already been demonstrated through the

validation of the Tempus HRD-RNA model, which was shown to predict *BRCA1/2* status in prostate cancer within a Clinical Laboratory Improvement Amendments-certified clinical setting (Leibowitz et al. 2022). However, there remains a need for clinically relevant models for predicting HR deficiency in breast cancer.

Several other transcriptional signatures for HR deficiency have been developed, albeit not in a clinical setting. These include include a 70-gene chromosomal instability signature, CIN70, defined by copy number alterations characteristic of HR deficiency, a 77-gene *BRCA1ness* signature defined by expression patterns characteristic of *BRCA1*-impaired breast cancers, and a 7-gene PARP inhibitor sensitivity signature (PARPi7) defined by molecular features associated with response to olaparib in breast cancer (Carter et al. 2006; Severson et al. 2017; Daemen et al. 2012). While these signatures have proven adept at predicting HR deficiency and PARP inhibitor response, they do not individually provide a complete picture of the HRD phenotype (Jacobson et al. 2023; Severson et al. 2017; Daemen et al. 2012). CIN70 is constructed based solely on copy number aberration patterns, and therefore is not applicable to the full range of possible HR deficiency features in transcriptomic datasets (Carter et al. 2006). Severson *et al.*'s *BRCA1ness* signature describes only *BRCA1*-negative HRD expression, failing to address other types of HR deficiency which may exhibit intact *BRCA1* (Severson et al. 2017). Finally, PARPi7 and the *BRCA1ness* signature have been proven to be predictive of response to just one of many possible PARP inhibitors - PARPi7 of olaparib response and the *BRCA1ness* signature of veliparib + carboplatin response (Daemen et al. 2012; Severson et al. 2017).

A 230-gene transcriptional HR deficiency signature was developed by Peng *et al.* through microarray differential expression analysis for control and HRD human mammary epithelial cell lines (Peng et al. 2014). Genes were selected if their expression between the control and HRD

cells differed by a factor of 2 or more, and the resulting signature is predictive of HR deficiency and sensitivity to olaparib and rucaparib in breast cancer cells. More recently, a 228-gene transcriptional signature was developed by Jacobson *et al.* to characterize the transcriptomic profiles of HRD and HRP breast cancers (Jacobson et al. 2023). They performed a multinomial elastic net regression to select genes and then employed a nearest centroid method to develop unique templates for HR deficiency, HR proficiency, *BRCA1ness*, *BRCA2ness* and *BRCA*-positive HR deficiency. The final signature successfully classifies HRD and HRP samples and distinguishes between responders (R) and non-responders (NR) to rucaparib, niraparib, olaparib, and talazoparib in breast cancer. While these signatures aptly predict HR deficiency and PARP inhibitor response as well as address the shortcomings of CIN70, PARPi7 and the *BRCA1ness* signature, they require excessively large expression profiles for classification. Moreover, they are not corrected for breast cancer molecular subtypes. The basal-like breast cancer subtype has been linked to HR deficiency (Sorlie et al. 2003; Anders et al. 2010), and therefore these models may be predicting a specific molecular subtype rather than the HRD/HRP phenotypes.

Presented here is a 153-gene transcriptional signature for characterizing HR deficiency in breast cancer. Genes were selected and weighed through training and validation for a linear support vector classifier using bulk RNA-sequencing (RNA-Seq) data from The Cancer Genome Atlas's breast cancer cohort (TCGA-BRCA). The clinical significance of the resulting transcriptional signature was tested on independent transcriptomic data with PARP inhibitor-response annotations from the I-SPY2 clinical trial (Barretina et al. 2012; Pusztai et al. 2021). The signature is able to detect HR deficiency from transcriptomic data as well as identify breast cancer patients who may be sensitive to PARP inhibitors.

Chapter 2 Training and Validation of Machine Learning Models to Estimate Transcriptional Signatures of Homologous Recombination Deficiency in Breast Cancer

Machine learning (ML) is a useful technique for extracting patterns that are concealed within complex, high-dimensional datasets, such as gene expression datasets, which can be exceedingly large and difficult to interpret. Thus, an ML model was defined to extract a transcriptional signature of HR deficiency and/or PARP-inhibitor response from transcriptomic data from breast cancer patients.

There are two main categories of machine learning: unsupervised and supervised. In unsupervised ML, a model separates entries of an unlabeled dataset into clusters. Entries within the same cluster are predicted to be similar, and entries in adjacent clusters are predicted to be distinct, or less closely related. Since the data is unlabeled, the meanings of these clusters are not immediately evident. In supervised ML, a model is trained on a labeled dataset, whereby each entry already belongs to a known group or label. The model then learns which features in the dataset are most important for distinguishing between labels, and defines weightings for these features that optimize this separation. The transcriptional signature is intended to define an HRD and/or PARP-inhibitor sensitivity gene expression profile for breast cancer patients. Thus, the labels of interest are already known, and signature inference is a supervised machine learning problem.

Supervised ML model development is composed of three main steps: training, validation, and testing. The training process is the “learning” aspect of machine learning - the model evaluates a labeled “training” dataset to determine a transformation of the data that leads to optimal separation of distinctly labeled entries (*e.g.*, HRD versus HRP samples). This is also referred to as model fitting, as dataset features (*e.g.*, genes for transcriptomic data) are weighed

in proportion to their observed importance in this separation. Thus, features deemed important for separation are weighed higher, while features deemed less impactful are assigned lower weights. The set of weighted features that define a trained ML model can be interpreted as a signature that may be predictive of the training dataset labels. The predictive ability of these signatures can be investigated and optimized during validation and testing.

The validation step involves allowing the trained model to make predictions as to the underlying labels of each entry of a separate "validation" dataset - usually a held-out subset from the same source as the training data. The model's predictive performance is evaluated and its parameters are re-adjusted repetitively until satisfactory performance metrics are achieved. The final step, testing, usually requires a "test" dataset from a source that is independent from the training and validation data. The model makes predictions on this unseen data, and its performance is evaluated to determine the ability of the model to generalize beyond the specific data it has been trained on. This step is intended to simulate the efficacy and relevance of the model when it comes to making "real-world" predictions.

Following these steps, ML models were developed to estimate candidates for an HR deficiency transcriptional signature.

Preparation of transcriptomic data for training and validation

Raw and fragments per kilobase of transcript per million mapped reads (FPKM) bulk RNA-Sequencing expression profiles for 1,231 samples in TCGA-BRCA were obtained from the Genomics Data Common's Data Portal for model training and validation. Duplicate, normal tissue, and male patient samples were removed from the datasets, and only protein-coding genes were retained. Additionally, mitochondrial, ribosomal, and non-autosomal genes were removed,

as well as genes expressed in less than 5% of all samples. After filtering, the expression datasets each contained the same 1,100 samples and 17,979 protein-coding genes.

Samples were given HRD/HRP labels using HRD scores and *BRCA1/2* annotations for TCGA-BRCA obtained from genomics analyses in Steele *et al.* and Polak *et al.*, respectively (Steele et al. 2022; Polak et al. 2017). Specifically, samples were labeled HRD if they exhibited *BRCA1/2* loss and/or an HRD score ≥ 50, and HRP if they exhibited no *BRCA1/2* loss and an HRD score ≤ 11. *BRCA1/2* loss was defined as either epigenetic silencing or biallelic inactivation of *BRCA1* or *BRCA2*. While an HRD score cutoff of at least 42 has been found to be predictive of HR deficiency (Telli et al. 2016), HRD and HRP score cutoffs were instead chosen at the extreme ends of the TCGA-BRCA HRD score distribution, at 50 and 11 respectively (**Fig. 2.1**). This was done to ensure that the models were trained on unambiguous, high-confident HRD and HRP transcriptomic profiles, therefore encouraging the models to extract the most important features for HRD/HRP separation and discouraging the misinterpretation of noise or uninformative features in the data as relevant. Thus, samples with non-extreme HRD scores (*i.e.*, greater than 11 and less than 50) could not be labeled HRP, and were labeled HRD only if they also showed evidence of *BRCA1/2* loss. Samples for which HR labels were indeterminable by these guidelines were excluded from further analyses. Ultimately, a total of 433 samples remained, with 244 samples labeled as HRP and 189 samples labeled as HRD.

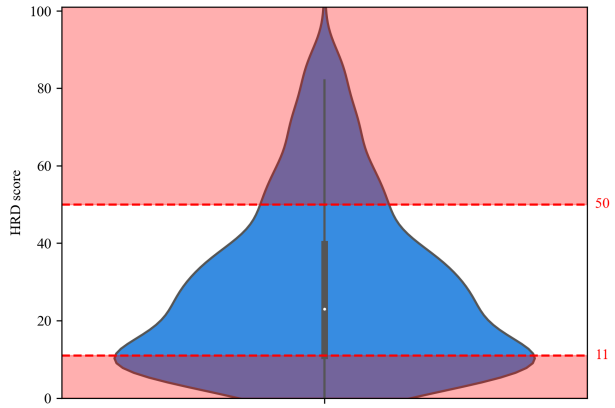

Figure 2.1: Distribution of HRD scores for samples in the TCGA-BRCA training and validation datasets.

PAM50 molecular subtype annotations from Berger et al. (Berger et al. 2018) revealed a significant imbalance in the subtype distributions of HRD and HRP samples in the data, such that the majority of HRD samples were defined as basal-like and the majority of HRP samples were defined as luminal A (**Fig. 2.2a**). This phenomenon has been observed previously, as have similarities between HR deficiency and the basal-like subtype in breast cancer (Anders et al. 2010; Livasy et al. 2006; Prat et al. 2014; Sorlie et al. 2003; Swain 2008; Jacobson et al. 2023). This raised the concern that a model trained to predict HR deficiency versus HR proficiency from this data may actually predict the basal-like versus luminal A subtype instead. To address this, the data was first randomly split into training and validation subsets, with 80% of samples (346/433) allocated for training and 20% (87/433) allocated for validation. HRD/HRP subtype imbalance was similarly observed in both the training and validation subsets (**Fig. 2.2b-c**). Samples were removed from the training data until the HRD and HRP subtype distributions were identical, and the removed samples were added to the validation data (**Fig. 2.2d-f**).

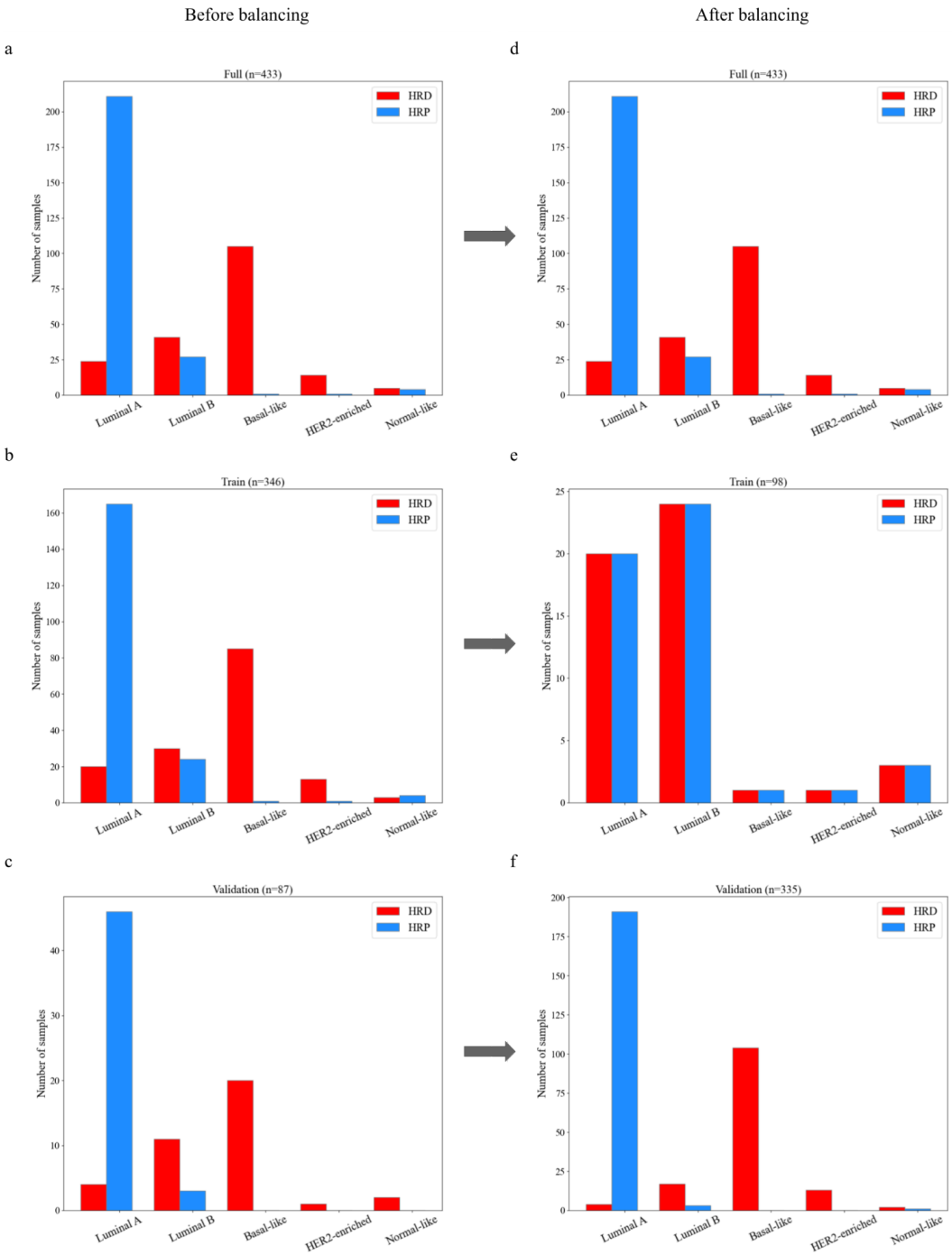

Figure 2.2: PAM50 molecular subtype distributions of HRD and HRP samples in: ***a)*** the full dataset, ***b)*** the training subset, and ***c)*** the validation subset before subtype balancing, and ***d)*** the full dataset, ***e)*** the training subset, and ***f)*** the validation subset after subtype balancing.

After filtering, labeling, and balancing the data for molecular subtype, training and validation of ML models could begin. The final sample sizes of the training and validation subsets allocated for these purposes were 98 and 335, respectively (**Fig. 2.3**).

Training data breakdown

	HRD	HRP
Luminal A	20	20
Luminal B	24	24
Basal-like	1	1
HER2-enriched	1	1
Normal-like	3	3
Total (HRD, HRP)	**49**	**49**
Total (HRD + HRP)		**98**

Validation data breakdown

	HRD	HRP
Luminal A	4	191
Luminal B	17	3
Basal-like	104	0
HER2-enriched	13	0
Normal-like	2	1
Total (HRD, HRP)	**140**	**195**
Total (HRD + HRP)		**335**

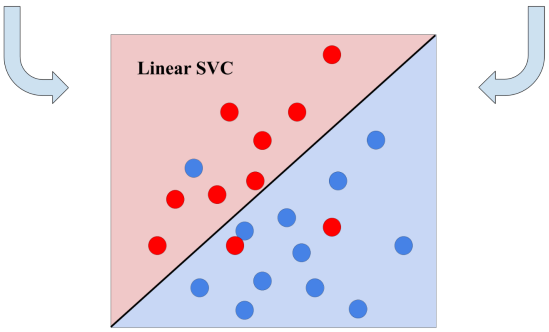

Figure 2.3: HRD/HRP and molecular subtype makeup of training and validation datasets.

Training and validation of linear support vector classifier models

Distinguishing between HRD and HRP gene expression profiles is an example of a binary classification problem in supervised machine learning, whereby an ML model is trained to assign new entities to one of two defined categories. Support vector classifiers (SVCs) are a class of ML models applicable to binary classification problems. For linearly separable data, SVCs define a hyperplane that optimally separates entities of different classes. For linearly inseparable data, SVCs define a kernel function to transform the data such that it becomes separable by a linear hyperplane. Linear SVCs employ linear kernel functions and therefore are optimal for use on linearly separable data. While neither linear separability, nor linear inseparability, of the

transcriptomic data was known, linear SVCs were utilized to separate HRD and HRP gene expression profiles due to their speed and efficacy on high-dimensional datasets with relatively few samples.

Three distinct linear SVC models arose from different feature selection methods and the use of either raw or FPKM-normalized TCGA-BRCA RNA-Sequencing data as input. Both differential expression (DE) analyses and L1 regularization were performed for feature selection purposes. Differential expression analyses highlight genes which exhibit extreme, significant differences in their expression by defined experimental groups. On the other hand, L1 regularization is an ML technique that drives the weights of uninformative features in a dataset to zero, thus defining a reduced set of high-relevance features. This is in contrast to L2 regularization, a related technique which drives the weights of uninformative features close to zero but does not allow them to reach zero exactly. For this reason, L2 regularization is not useful for feature selection; however, it is useful for model fitting, as it permits heavier weighting of informative features and encourages lower weighting of less-informative features. Before training and validation, all data were subjected to log2 transformation, and each expression value was centered within the datasets to the 75th quantile of the associated gene. Fixed values of 9.5 were also added to each expression value to avoid negative values in the data.

The first linear SVC model, M1, was trained and validated on FPKM-normalized data. A preliminary linear SVC was fitted on the FPKM training dataset to select features for M1 using L1 regularization. Of the 17,979 genes in the dataset, 83 were selected as features, and the training and validation feature space was reduced accordingly. M1 was fit to, or trained on, the reduced training dataset using L2 regularization, through which feature weights were defined (**Fig. 2.4**).

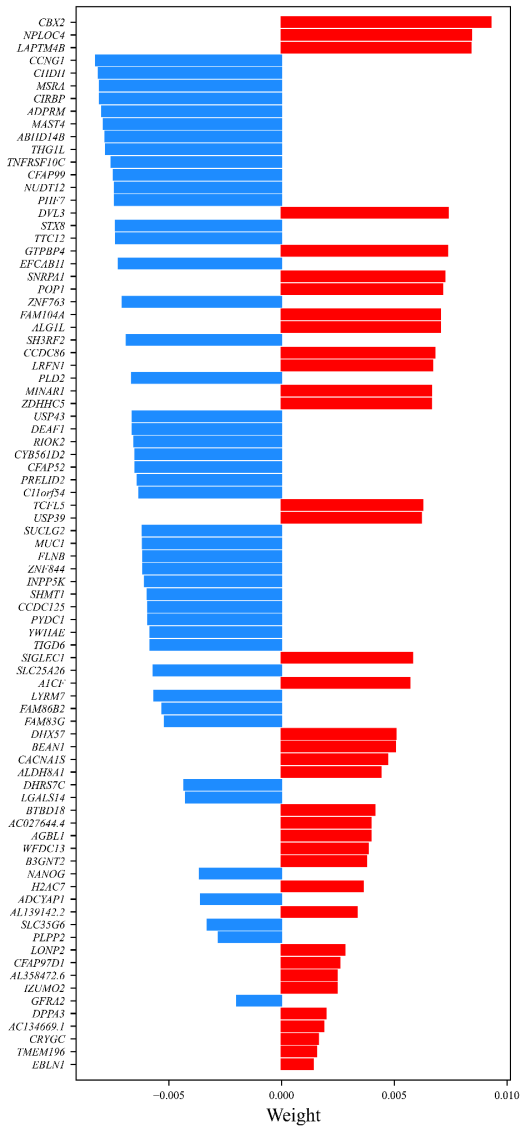

Figure 2.4: Summary of M1 feature importance. Features are ranked from most important (top) to least important (bottom). Negative weights indicate a feature predictive of HR proficiency; positive weights indicate a feature predictive of HR deficiency.

Functional enrichment analyses were performed on the features using g:Profiler's g:GOSt module (Kolberg et al. 2023) (**Fig. 2.5**). Top results for HRD-predictive features include DNA demethylation, misfolded protein clearing, mis-splicing, and inhibition of cell proliferation, metastasis and tumor development; top results for HRP-predictive features include tumor growth, metastasis suppression, lipid metabolism, metabolism of glycine, serine, and threonine,

and the methionine salvage pathway. While many of the top results for both HRD- and HRP-predictive features are relevant in the context of cancer, none were statistically significant (padj > 0.05). This result undermines the credibility of M1, as the features are expected to be significantly enriched for functional attributes related to the HRD and/or HRP phenotype. However, despite insignificant functional enrichment results, M1 demonstrated strong performance in HRD/HRP prediction on the held-out validation dataset with an area under the receiver operating characteristic curve (AUC) of 0.96 (**Fig 2.6**).

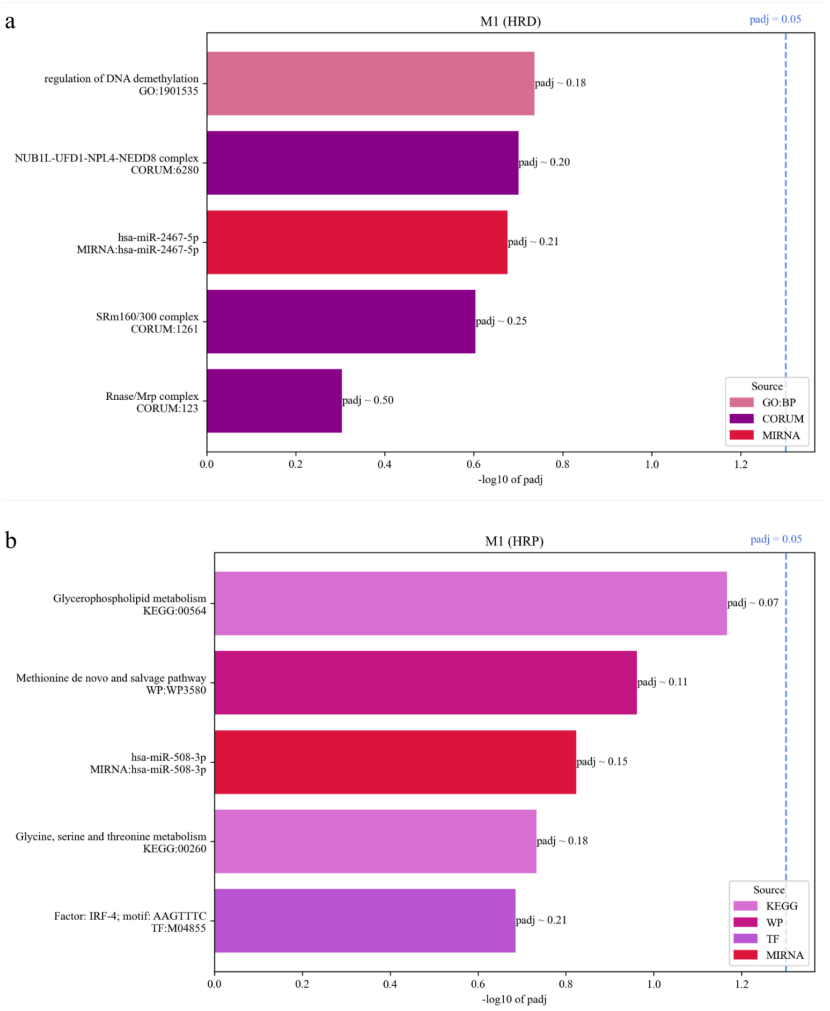

Figure 2.5: Top 5 functional enrichment results for ***a)*** HRD-predictive features and ***b)*** HRP-predictive features of M1. P-values are measured using a Fisher's one-tailed test, and corrected using g:Profiler's g:SCS method (Kolberg et al. 2023).

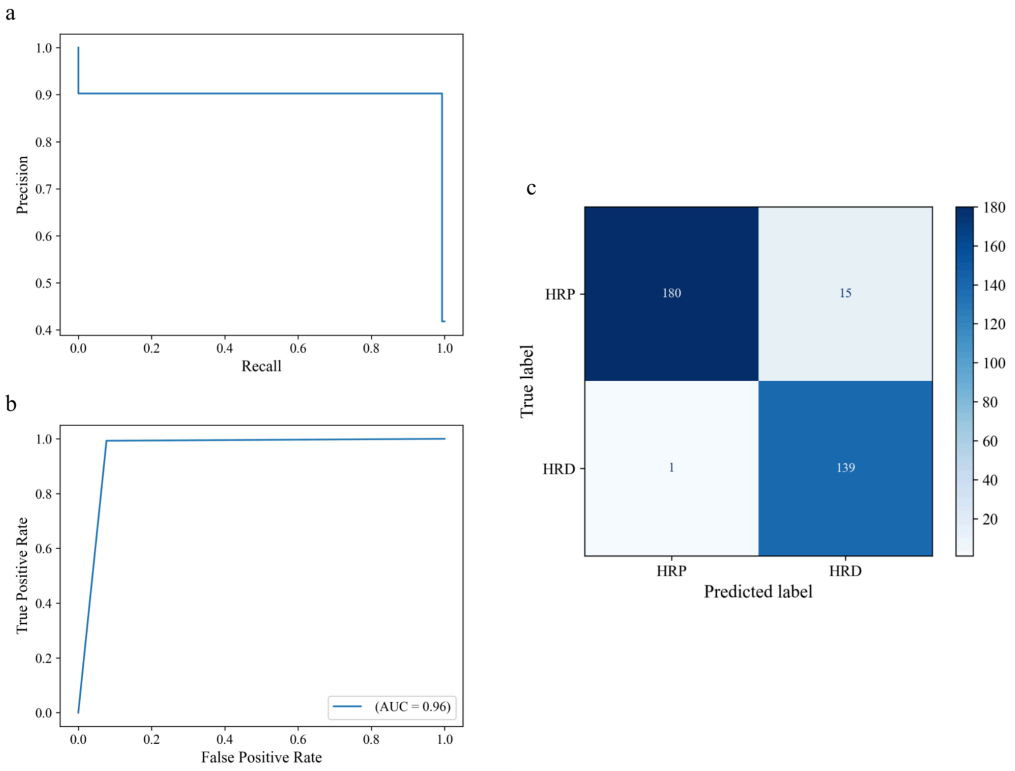

Figure 2.6: Performance metrics for M1 on the FPKM validation dataset. ***a)*** Precision-recall curve. Precision is defined as the ratio of true HRD predictions to total HRD predictions; recall is defined as the ratio of true HRD predictions to total HRD samples. ***b)*** Receiver operating characteristic (ROC) curve. The true positive rate is the probability that M1 will correctly classify HRD samples as HRD; the false positive rate is the probability that M1 will falsely classify HRP samples as HRD. An area under the ROC curve (AUC) value of 1 indicates perfect classification, while an AUC value of 0.5 indicates the model has no discriminatory power. ***c)*** Confusion matrix displaying the number of correct and incorrect HRD and HRP predictions.

The second model, M2, was trained and validated on the raw dataset. L1 regularization resulted in the selection of 88 features, with feature weights determined through model training and L2 regularization (**Fig. 2.7**).

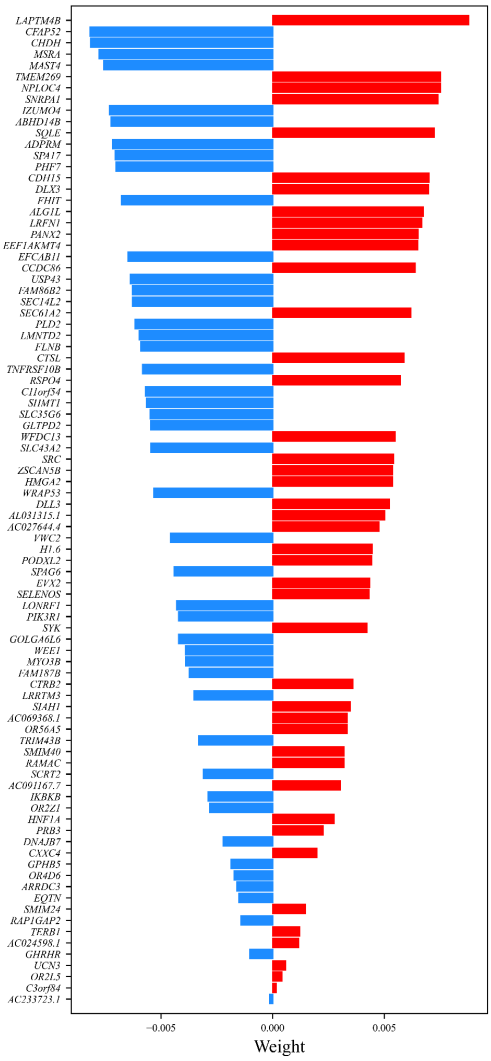

Figure 2.7: Summary of M2 feature importance.

M2 derived more significant results from functional enrichment analysis compared to M1 (**Fig. 2.8**). Top functional enrichment terms for HRD-predictive features are related to cell migration and proliferation, misfolded protein clearing and degradation, and immune and inflammatory response; top terms for HRP-predictive features are related to apoptosis, metastasis, cell proliferation, and pancreatic and lung cancer - though, notably, not breast cancer. While several significant results were found for HRP-predictive features, thus marking an

improvement upon M1, still only one statistically significant result was found for HRD-predictive features.

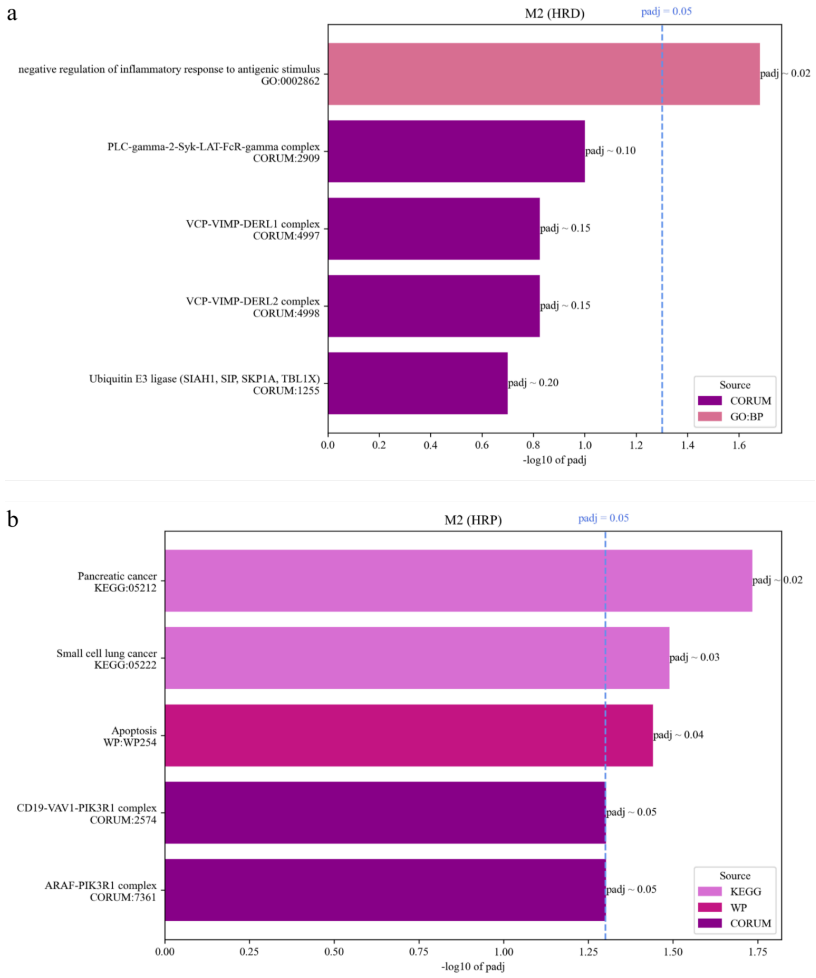

Figure 2.8: Top 5 functional enrichment results for ***a)*** HRD-predictive features and ***b)*** HRP-predictive features of M2.

M2's validation performance metrics were very similar to those of M1, with a slightly reduced AUC of 0.95 (**Fig. 2.9**).

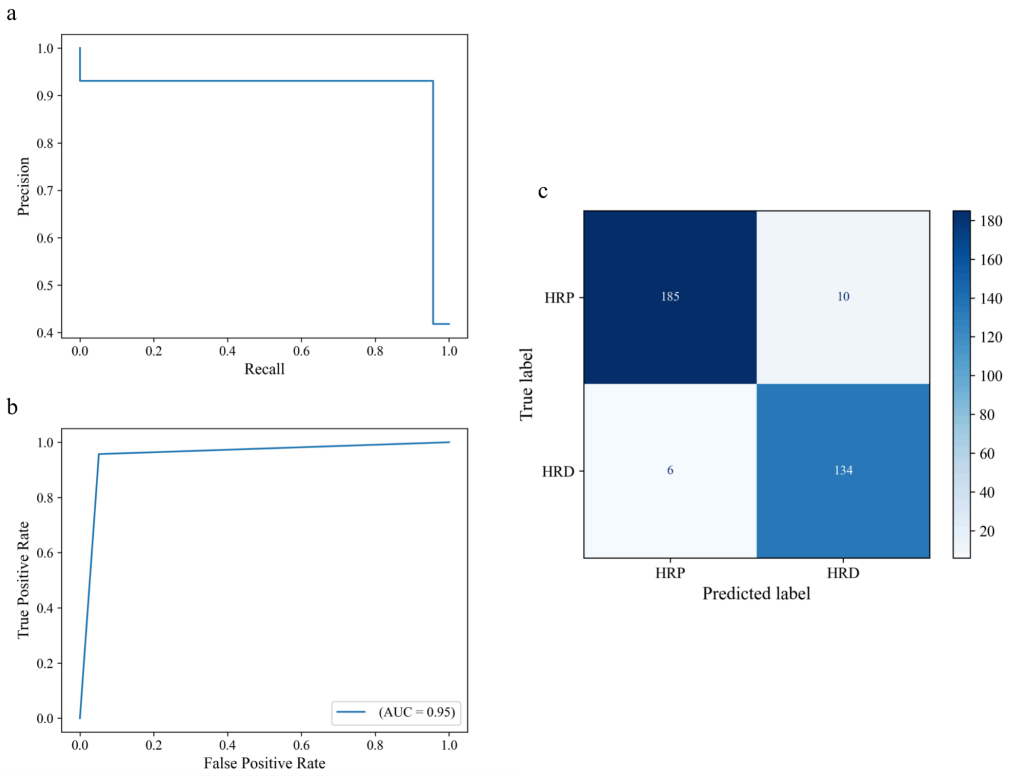

Figure 2.9: Performance metrics for M2 on the raw validation dataset. ***a)*** Precision-recall curve. ***b)*** ROC curve. ***c)*** Confusion matrix.

The third and final model, M3, was trained and validated on the raw dataset, with DE analysis performed for feature selection using the DESeq2 package (Love et al. 2014). While, for M1 and M2, log2 transformation and centering occurred before feature selection, for M3 these transformations occurred after feature selection by DESeq2, citing the need for un-normalized and un-transformed DESeq2 input (Love et al. 2014).

DESeq2 associates each gene with a log2 fold change value as well as an adjusted p-value. The log2 fold change values describe the extent to which each gene is differentially expressed in HRP versus HRD samples in the training dataset. For example, a log2 fold change value of -1 for a particular gene means that the expression level of that gene in HRP samples is half the expression level of the gene in HRD samples, as 2^{-1}=0.5. A log2 fold change value of 1,

on the other hand, indicates that the expression level of the gene in HRP samples is twice the expression level of the gene in HRD samples, as $2^1=2$. Therefore, negative log2 fold change values indicate a gene which is predictive of HR deficiency, while positive log2 fold change values indicate a gene which is predictive of HR proficiency. P-values were calculated by the Wald test and corrected according to the Benjamini Hochberg method. 153 genes, with associated padj < 0.01 and |log2 fold change| > 2, were selected as features for M3 (**Fig. 2.10**). L2 regularization was employed during fitting, with the resulting feature weights summarized in **Figure 2.11**.

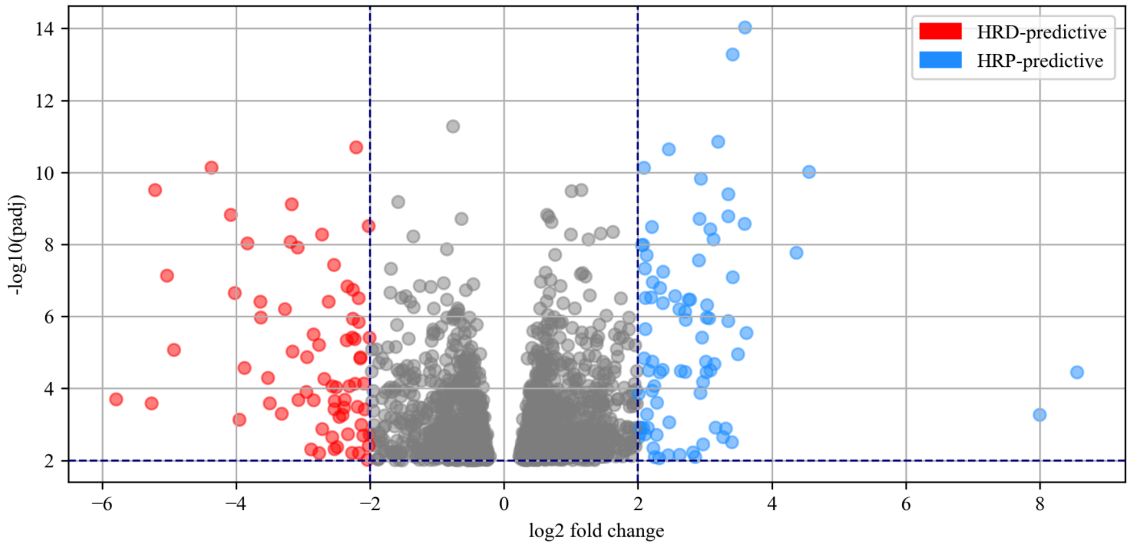

Figure 2.10: Adjusted p-value and log2 fold change thresholds applied for feature selection during DE analysis.

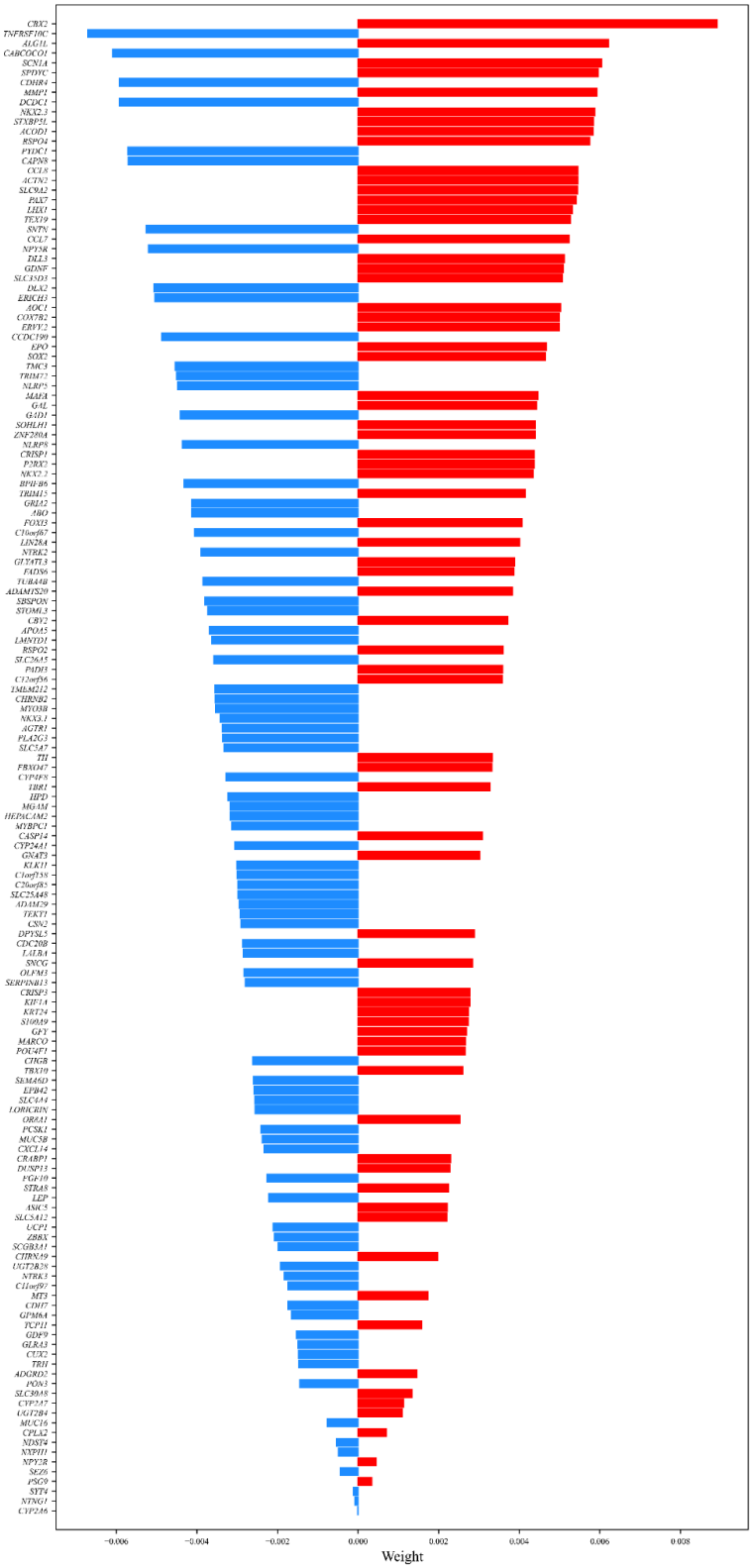

Figure 2.11: Summary of M3 feature importance.

Functional enrichment results for M3 were the most significant of all three models (**Fig. 2.12**). Top functional enrichment terms for HRD-predictive features were related to cell proliferation, metastasis inhibition, cancer stem cell formation, cancer therapeutic resistance, inflammation, and DNA damage response regulation; top terms for HRP-predictive features were related to tumor growth, apoptosis, angiogenesis, metastasis, cell motility, cancer-cell invasion, cytokinesis, and lactating breast ductal cells. M3 also achieved very similar performance metrics to M1 and M2 during validation, with an AUC of 0.95 (**Fig. 2.13**).

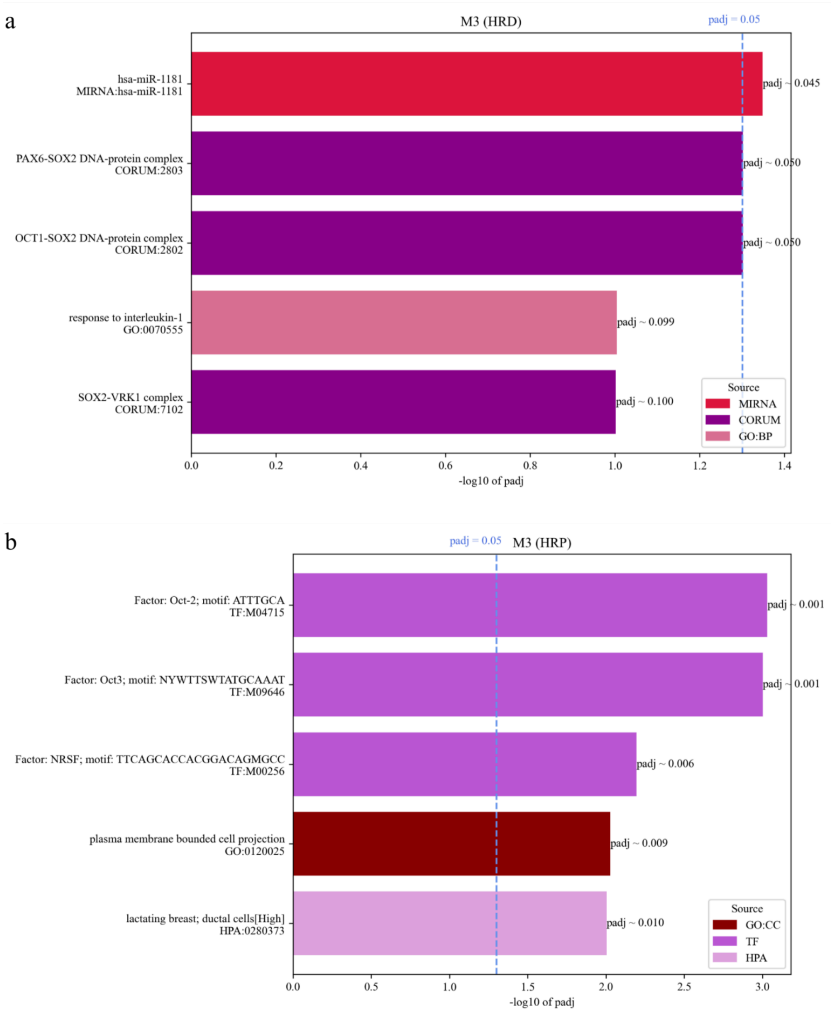

Figure 2.12: Top 5 functional enrichment results for ***a)*** HRD-predictive features and ***b)*** HRP-predictive features of M3.

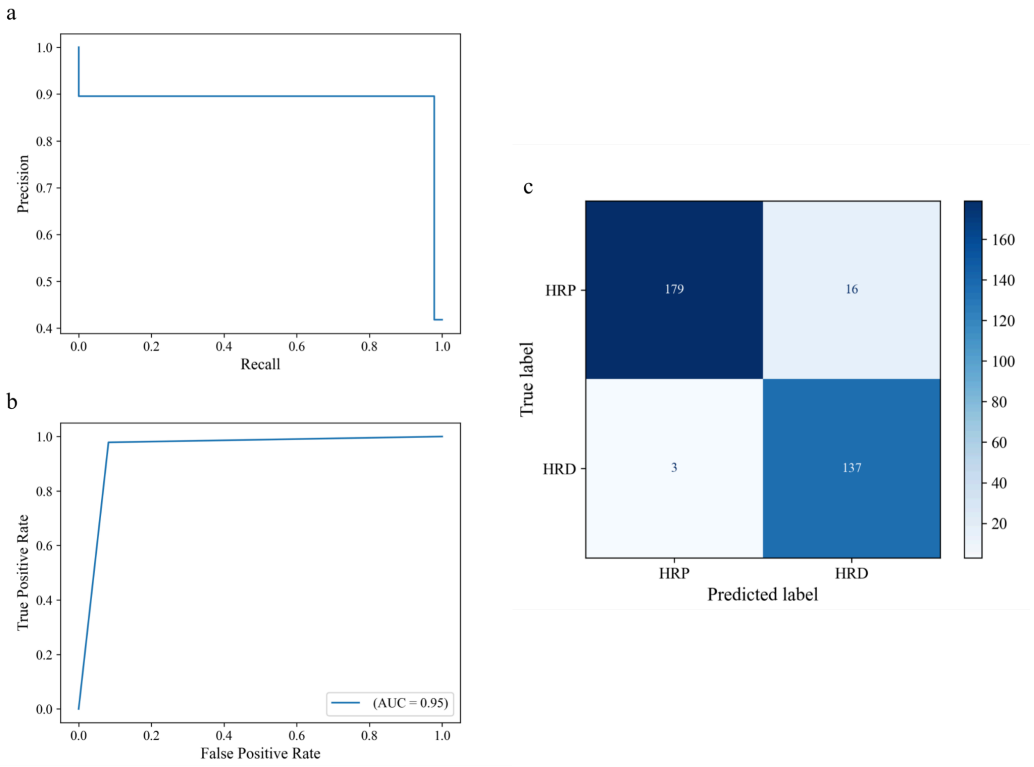

Figure 2.13: Performance metrics for M3 on the raw validation dataset. ***a)*** Precision-recall curve. ***b)*** ROC curve. ***c)*** Confusion matrix.

Feature overlap was minimal amongst M1, M2, and M3 (**Fig. 2.14**); nevertheless, validation performance was comparable across all three models (**Table 2.1**). Given that, of the three models, M3 exhibited a superior functional enrichment result, M3 was expected to perform optimally on the independent test dataset.

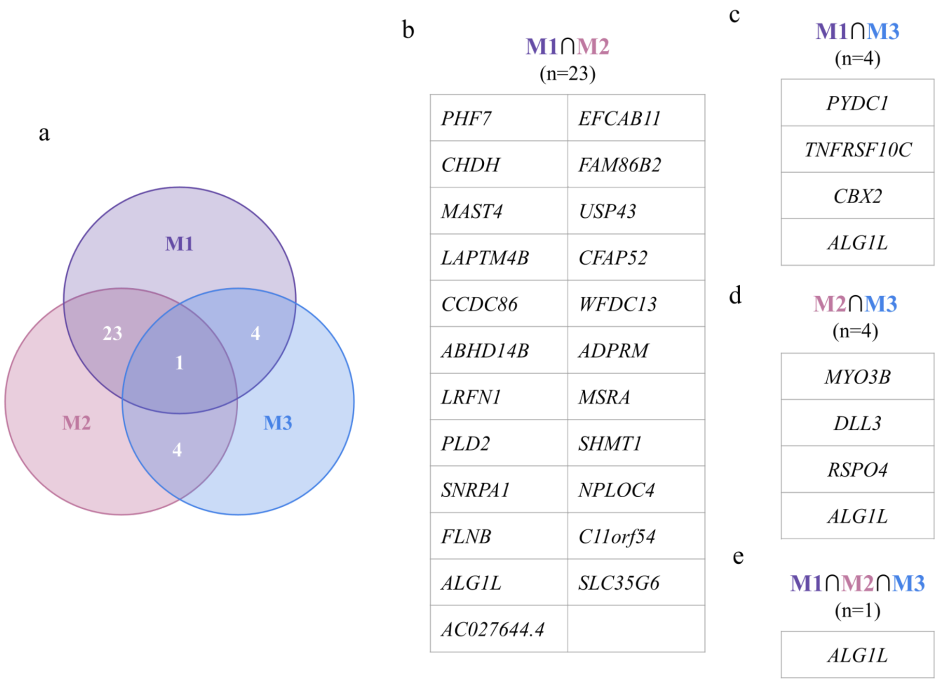

Figure 2.14: Summary of feature overlap. ***a)*** Venn diagram quantifying overlap amongst M1, M2 and M3 features. ***b)*** Features common to both M1 and M2. ***c)*** Features common to both M1 and M3. ***d)*** Features common to both M2 and M3. ***e)*** Features common to M1, M2 and M3.

Table 2.1: Summary of model attributes and validation performance metrics.

	M1	**M2**	**M3**
Input data	FPKM RNA-Seq	Raw RNA-Seq	Raw RNA-Seq
Scaling method	log2 + centering to 75th quantile	log2 + centering to 75th quantile	log2 + centering to 75th quantile
Feature selection	L1 regularization	L1 regularization	DE analysis
Number of features	83	88	153
Precision	0.90	0.93	0.90
Recall	0.99	0.96	0.98
AUC	0.96	0.95	0.95

Alternative strategies and potential pitfalls

There are several opportunities for further exploration and potential improvement in the training and validation process. Alternative methods for data processing may include transcripts per million (TPM) and/or trimmed mean of M-values (TMM) normalization, as well as z-score and MinMax scaling.

Furthermore, while the models are trained to predict HR deficiency from gene expression data, which describes only the most recent status of the HR pathway, the ground-truth definitions for HR deficiency and HR proficiency supplied for training are derived from genomic data, and thus are still historic markers. Ideal ground-truth definitions for training should describe PARP inhibitor and/or platinum therapy response, as these are the clinical consequences of HR deficiency; thus, defining ground-truth using clinical datasets would be an optimal approach for future analyses.

Various SVC kernel functions can also be explored. A linear kernel was chosen due to its speed on high-dimensional data, and while it demonstrated efficacy in the training and validation steps, linear separability cannot be assumed of the data. Thus, it is possible that other kernel functions, such as polynomial, radial basis, or sigmoid functions, may instead be optimal for distinguishing HRD from HRP transcriptomic profiles. Cross validation could be implemented in the future to determine the optimal kernel function for HRD/HRP classification.

Alternative ML models can also be considered for HRD/HRP transcriptional signature estimation. Both random forest and gradient-boosting classifiers define an optimal set of features and weights for object classification that can be interpreted as a class-predictive signature. Furthermore, it is possible to implement regression for HRD/HRP differentiation, whereby samples would be assigned a value indicating their position on a spectrum between HRD and

HRP rather than just an HRD/HRP label. Regressors that may be applicable for transcriptional signature inference include lasso, elastic net and support vector regressors.

Lastly, it would be informative to evaluate model performance on samples from different cancer types that have been linked to HR deficiency, such as ovarian, pancreatic, and prostate cancers (Stewart et al. 2022).

Chapter 3 Selecting an Optimal Transcriptional Signature for Homologous Recombination Deficiency in Breast Cancer Through Evaluation of Independent Testing Performance

All three models displayed robust validation performance metrics (**Table 2.1**). However, the training and validation data were both generated as subsets of the same TCGA-BRCA dataset; thus, evaluation of the models on an independent test dataset was necessary to determine whether they exhibited robust performance in general HRD/HRP prediction, rather than merely overfitting the TCGA-BRCA data. Overfitting is a common machine learning issue where models learn the cohort-specific patterns that are only found in the training dataset along with the relevant features. An overfitted model often performs extremely well during validation, but fails to generalize to unseen data during testing. Thus, the three models were tested for their ability to generalize to previously unseen data, which were generated independently from TCGA-BRCA. Additionally, data with drug response annotations were selected for independent testing to evaluate the clinical relevance of the transcriptional signatures generated by the models.

Preparation of independent test dataset

Microarray and clinical data for the I-SPY2 clinical trial (Pusztai et al. 2021) were obtained from the Gene Expression Omnibus database under the accession number GSE173839. The I-SPY2 dataset contains log2-normalized microarray signals for 21,508 genes for 105 breast cancer patients, 34 from the control group and 71 from the experimental treatment group. The experimental group was treated with the chemotherapy drug taxol, along with durvalumab, a PD-L1 inhibitor, and the PARP-inhibitor olaparib, while the control group was treated only with taxol. These treatments were followed by doxorubicin/cyclophosphamide in both groups.

Pathological complete response (pCR) annotations were used in place of HRD/HRP labels, where a pathological complete response is defined as the absence of all cancer cells in a tissue sample following the experimental treatment regimen. As such, a complete response (pCR=1) corresponds with clinical HR deficiency, while a failed complete response (pCR=0) corresponds with clinical HR proficiency. In the experimental group, 29 out of 71 patients were responders (R), or achieved a pathological complete response, while 42 out of 71 patients were non-responders (NR), or had a failed complete response. Patients in the control group were not annotated for PARP-inhibitor pCR and were consequently excluded from further analysis. Additionally, genes with missing expression values were removed from the dataset. The trained models were applied to the final I-SPY2 dataset consisting of expression profiles for 71 patients across 21,434 genes.

Independent testing summary for linear SVC models

Only 72 of M1's 83 features were present in the test dataset for R/NR prediction. In stark contrast with its validation performance, M1 performed poorly on the test dataset, classifying every sample as HRP/NR (**Fig. 3.1**).

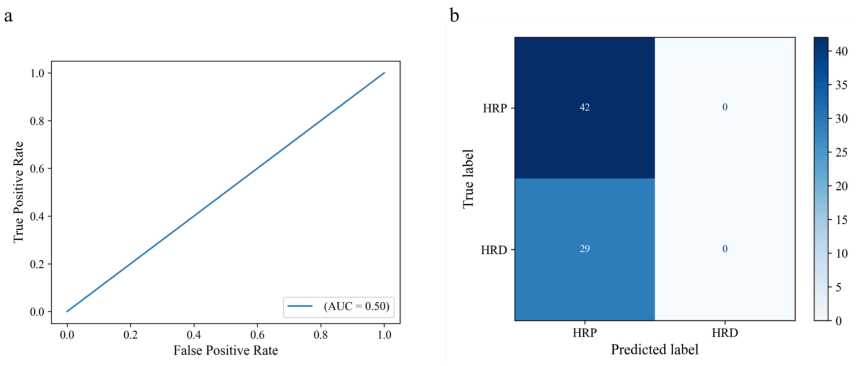

Figure 3.1: Performance metrics for M1 on the I-SPY2 test dataset; precision-recall curve is not shown due to undefined precision for M1. ***a)*** ROC curve. ***b)*** Confusion matrix.

M2's test performance was similarly poor, although all samples were classified as HRD/R instead (**Fig. 3.2**). Only 71 of M2's 88 features were present in the test dataset.

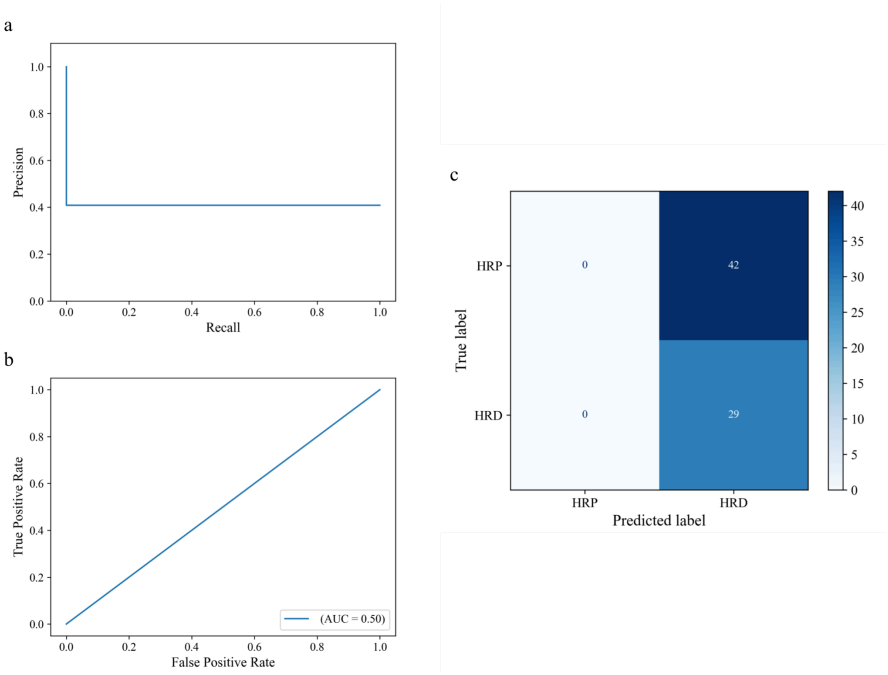

Figure 3.2: Performance metrics for M2 on the I-SPY2 test dataset. ***a)*** Precision-recall curve. ***b)*** ROC curve. ***c)*** Confusion matrix.

M3 was the only model with better-than-random prediction on the test dataset (**Fig. 3.3**). However, its predictive ability was heavily imbalanced between HRD/R and HRP/NR samples, with a steep false negative rate (13/29=0.45) leading to significant misclassification of HRD/R samples (**Fig. 3.3c**). 141 of M3's 153 features were retained in the test dataset.

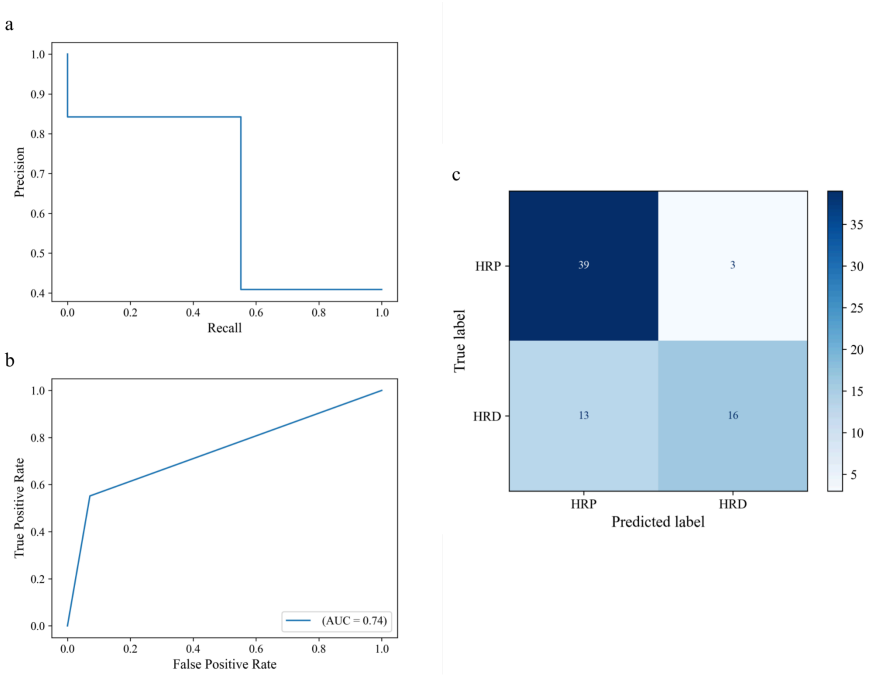

Figure 3.3: Performance metrics for M3 on the I-SPY2 test dataset. ***a)*** Precision-recall curve. ***b)*** ROC curve. ***c)*** Confusion matrix.

As hypothesized, M3 exhibited the best performance on the test dataset of all three models (**Table 3.1**). While its test efficacy is still diminished relative to its validation performance, it is important to note that the test dataset has been derived from a microarray, which is an entirely different gene expression quantification technique compared to RNA-Sequencing. Although M3 does not demonstrate ideal prediction on the test dataset, it does generalize well for data obtained from a completely different assay to its training dataset. Therefore, of the transcriptional signatures generated by each model, the signature produced by M3 is the optimal choice for detecting HR deficiency in the breast cancer transcriptome.

Table 3.1: Updated model summary to include test performance metrics. M3 outperforms M1 and M2 on the independent test dataset.

	M1	**M2**	**M3**
Input data	FPKM RNA-Seq	Raw RNA-Seq	Raw RNA-Seq
Scaling method	log2 + centering to 75th quantile	log2 + centering to 75th quantile	log2 + centering to 75th quantile
Feature selection	L1 regularization	L1 regularization	DE analysis
Number of features	83	88	153
Number of features in test dataset	72	71	141
Validation precision	0.90	0.93	0.90
Validation recall	0.99	0.96	0.98
Validation AUC	0.96	0.95	0.95
Test precision	0.00	0.00	0.84
Test recall	0.00	0.00	0.55
Test AUC	0.50	0.50	0.74

Importantly, the HR deficiency signature generated by M3 demonstrates superior performance compared to Jacobson *et al.*'s 228-gene signature in distinguishing responders to durvalumab and olaparib from non-responders in the I-SPY2 trial **(Fig. 3.4)** (Jacobson et al. 2023).

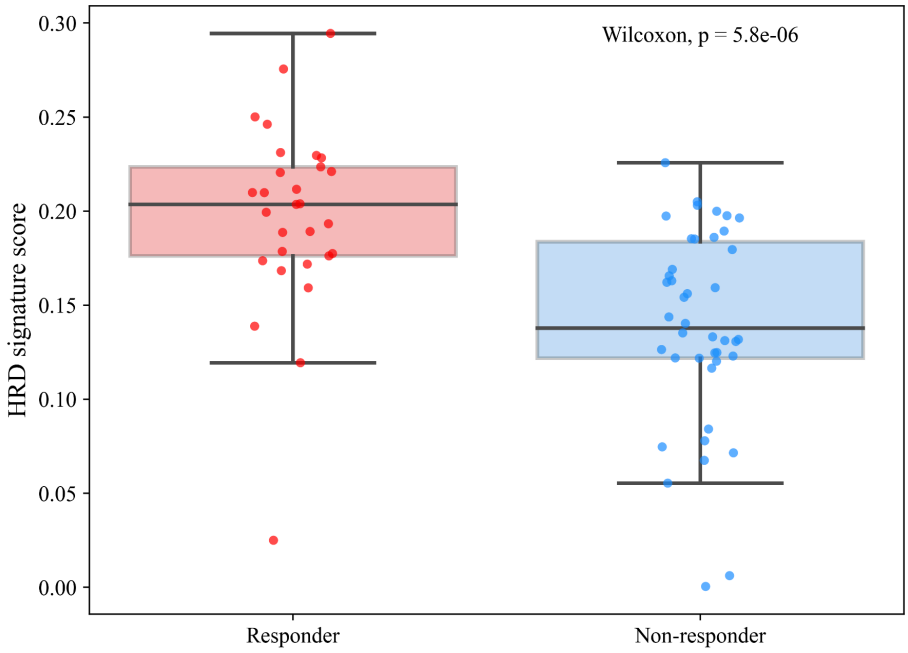

Figure 3.4: Comparison of HRD scores calculated by M3 for responders and non-responders to durvalumab and olaparib in the I-SPY2 trial.

This marks an improvement not only in identifying patients who may respond to PARP inhibitors but also in reducing the number of features required for prediction.

Alternative strategies and potential pitfalls

While M3 demonstrated generalizability on the test dataset, its performance was not optimal. This may reflect the difference in ground-truth definitions used for training and testing, as M3 was trained using genomic HRD/HRP definitions but tested against clinical response to PARP inhibitors. M3's test performance may also have been impacted by the differences of microarray technology for gene expression quantification as compared to RNA-Sequencing. For example, the reliance of microarray results on probe hybridization limits the detection of RNA transcript variants and can lead to increased noise due to non-specific binding of probes and target transcripts (Wang et al. 2009; Okoniewski and Miller 2006). The range of detection of

microarrays is limited by noise at lower expression levels and, at higher expression levels, the inability to differentially quantify expression past a maximum detection capacity (Wang et al. 2009).

The I-SPY2 dataset also lacked data for 12 of M3's 153 features. Several of these features were identified as highly important for HRD/HRP classification, and their absence may have contributed to the suboptimal performance of M3 on the test dataset. Additionally, the log2 fold change cutoffs of 2 and -2 used for feature selection for M3 are stringent, preventing genes with less than four-fold expression differences between HRD and HRP samples in the training dataset from being selected as features. Test dataset performance may be further improved by including more features for prediction through the use of less stringent log2 fold change cutoffs. Future directions include extending the transcriptional signature to additional HR deficiency-associated cancers (*e.g.*, ovarian, pancreatic, and prostate cancers) (Stewart et al. 2022) and evaluating its efficacy in predicting responses to PARP inhibitors beyond olaparib.

CONCLUSION

In light of a breast cancer diagnosis, it is essential that the prescribed treatment plan offers the greatest possible improvement of survival. Furthermore, due to the time-sensitive nature of the disease, an optimal treatment plan must be selected and initiated as soon as possible. For one, a timely diagnosis of HR deficiency in breast cancer allows for the immediate initiation of PARP inhibitors and/or platinum-based therapy, significantly improving prognoses. While several tools have been developed for HR deficiency prediction, the signature presented here is unique in its applicability and efficacy. It leverages transcriptomic data to identify HR deficiency based on only the most recent observed phenotype rather than the mutational history of the cell. The signature is also widely-applicable, focusing on transcriptional patterns common to all breast cancer subtypes and mechanisms of HR deficiency. Moreover, the signature allows identification of breast cancer patients who may be sensitive to PARP inhibitors based on transcriptomic data.

REFERENCES

Abbas-Aghababazadeh, F., Li, Q., & Fridley, B. L. (2018). Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. *PloS One*, *13*(10), e0206312. https://doi.org/10.1371/journal.pone.0206312

Abkevich, V., Timms, K. M., Hennessy, B. T., Potter, J., Carey, M. S., Meyer, L. A., Smith-McCune, K., Broaddus, R., Lu, K. H., Chen, J., Tran, T. V., Williams, D., Iliev, D., Jammulapati, S., FitzGerald, L. M., Krivak, T., DeLoia, J. A., Gutin, A., Mills, G. B., & Lanchbury, J. S. (2012). Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *British Journal of Cancer*, *107*(10), 1776–1782. https://doi.org/10.1038/bjc.2012.451

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., PCAWG Mutational Signatures Working Group, … PCAWG Consortium. (2020). The repertoire of mutational signatures in human cancer. *Nature*, *578*(7793), 94–101. https://doi.org/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., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–421. https://doi.org/10.1038/nature12477

Alhmoud, J. F., Woolley, J. F., Al Moustafa, A.-E., & Malki, M. I. (2020). DNA Damage/Repair Management in Cancers. *Cancers*, *12*(4). https://doi.org/10.3390/cancers12041050

Anders, C. K., Winer, E. P., Ford, J. M., Dent, R., Silver, D. P., Sledge, G. W., & Carey, L. A. (2010a). Poly(ADP-Ribose) polymerase inhibition: “targeted” therapy for triple-negative breast cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *16*(19), 4702–4710. https://doi.org/10.1158/1078-0432.CCR-10-0939

Anders, C. K., Winer, E. P., Ford, J. M., Dent, R., Silver, D. P., Sledge, G. W., & Carey, L. A. (2010b). Poly(ADP-Ribose) polymerase inhibition: “targeted” therapy for triple-negative breast cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *16*(19), 4702–4710. https://doi.org/10.1158/1078-0432.CCR-10-0939

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., … Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, *483*(7391), 603–607. https://doi.org/10.1038/nature11003

Berger, A. C., Korkut, A., Kanchi, R. S., Hegde, A. M., Lenoir, W., Liu, W., Liu, Y., Fan, H., Shen, H., Ravikumar, V., Rao, A., Schultz, A., Li, X., Sumazin, P., Williams, C., Mestdagh, P., Gunaratne, P. H., Yau, C., Bowlby, R., … Akbani, R. (2018). A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell*, *33*(4), 690–705.e9. https://doi.org/10.1016/j.ccell.2018.03.014

Birkbak, N. J., Wang, Z. C., Kim, J.-Y., Eklund, A. C., Li, Q., Tian, R., Bowman-Colin, C., Li, Y., Greene-Colozzi, A., Iglehart, J. D., Tung, N., Ryan, P. D., Garber, J. E., Silver, D. P., Szallasi, Z., & Richardson, A. L. (2012). Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discovery*, *2*(4), 366–375. https://doi.org/10.1158/2159-8290.CD-11-0206

Caldecott, K. W. (2008). Single-strand break repair and genetic disease. *Nature Reviews. Genetics*, *9*(8), 619–631. https://doi.org/10.1038/nrg2380

Carter, S. L., Eklund, A. C., Kohane, I. S., Harris, L. N., & Szallasi, Z. (2006). A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nature Genetics*, *38*(9), 1043–1048. https://doi.org/10.1038/ng1861

Carusillo, A., & Mussolino, C. (2020). DNA Damage: From Threat to Treatment. *Cells* , *9*(7). https://doi.org/10.3390/cells9071665

Ceccaldi, R., Rondinelli, B., & D'Andrea, A. D. (2016). Repair Pathway Choices and Consequences at the Double-Strand Break. *Trends in Cell Biology*, *26*(1), 52–64. https://doi.org/10.1016/j.tcb.2015.07.009

Corsello, S. M., Nagari, R. T., Spangler, R. D., Rossen, J., Kocak, M., Bryan, J. G., Humeidi, R., Peck, D., Wu, X., Tang, A. A., Wang, V. M., Bender, S. A., Lemire, E., Narayan, R., Montgomery, P., Ben-David, U., Garvie, C. W., Chen, Y., Rees, M. G., … Golub, T. R. (2020). Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*, *1*(2), 235–248. https://doi.org/10.1038/s43018-019-0018-6

Cortesi, L., Rugo, H. S., & Jackisch, C. (2021). An Overview of PARP Inhibitors for the Treatment of Breast Cancer. *Targeted Oncology*, *16*(3), 255–282. https://doi.org/10.1007/s11523-021-00796-4

Curtin, N. J., & Szabo, C. (2020). Poly(ADP-ribose) polymerase inhibition: past, present and future. *Nature Reviews. Drug Discovery*, *19*(10), 711–736. https://doi.org/10.1038/s41573-020-0076-6

Daemen, A., Wolf, D. M., Korkola, J. E., Griffith, O. L., Frankum, J. R., Brough, R., Jakkula, L. R., Wang, N. J., Natrajan, R., Reis-Filho, J. S., Lord, C. J., Ashworth, A., Spellman, P. T., Gray, J. W., & van't Veer, L. J. (2012). Cross-platform pathway-based analysis identifies markers of response to the PARP inhibitor olaparib. *Breast Cancer Research and Treatment*, *135*(2), 505–517. https://doi.org/10.1007/s10549-012-2188-0

Deans, A. J., & West, S. C. (2011). DNA interstrand crosslink repair and cancer. *Nature Reviews. Cancer*, *11*(7), 467–480. https://doi.org/10.1038/nrc3088

den Brok, W. D., Schrader, K. A., Sun, S., Tinker, A. V., Zhao, E. Y., Aparicio, S., & Gelmon, K. A. (2017). Homologous Recombination Deficiency in Breast Cancer: A Clinical Review. *JCO Precision Oncology*, *1*, 1–13. https://doi.org/10.1200/PO.16.00031

Dias, M. P., Moser, S. C., Ganesan, S., & Jonkers, J. (2021). Understanding and overcoming resistance to PARP inhibitors in cancer therapy. *Nature Reviews. Clinical Oncology*, *18*(12), 773–791. https://doi.org/10.1038/s41571-021-00532-x

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* , *21*(16), 3439–3440. https://doi.org/10.1093/bioinformatics/bti525

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, *4*(8), 1184–1191. https://doi.org/10.1038/nprot.2009.97

Ganai, R. A., & Johansson, E. (2016). DNA Replication-A Matter of Fidelity. *Molecular Cell*, *62*(5), 745–755. https://doi.org/10.1016/j.molcel.2016.05.003

Heeke, A. L., Pishvaian, M. J., Lynce, F., Xiu, J., Brody, J. R., Chen, W.-J., Baker, T. M., Marshall, J. L., & Isaacs, C. (2018). Prevalence of Homologous Recombination-Related Gene Mutations Across Multiple Cancer Types. *JCO Precision Oncology*, *2018*. https://doi.org/10.1200/PO.17.00286

Huang, L. C., Clarkin, K. C., & Wahl, G. M. (1996). Sensitivity and selectivity of the DNA damage sensor responsible for activating p53-dependent G1 arrest. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(10), 4827–4832. https://doi.org/10.1073/pnas.93.10.4827

Jackson, S. P., & Bartek, J. (2009). The DNA-damage response in human biology and disease. *Nature*, *461*(7267), 1071–1078. https://doi.org/10.1038/nature08467

Jacobson, D. H., Pan, S., Fisher, J., & Secrier, M. (2023). Multi-scale characterisation of homologous recombination deficiency in breast cancer. In *bioRxiv* (p. 2023.08.23.554414). https://doi.org/10.1101/2023.08.23.554414

Khanna, K. K., & Jackson, S. P. (2001). DNA double-strand breaks: signaling, repair and the cancer connection. *Nature Genetics*, *27*(3), 247–254. https://doi.org/10.1038/85798

Kolberg, L., Raudvere, U., Kuzmin, I., Adler, P., Vilo, J., & Peterson, H. (2023). g:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Research*, *51*(W1), W207–W212. https://doi.org/10.1093/nar/gkad347

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

Leibowitz, B. D., Dougherty, B. V., Bell, J. S. K., Kapilivsky, J., Michuda, J., Sedgewick, A. J., Munson, W. A., Chandra, T. A., Dry, J. R., Beaubier, N., Igartua, C., & Taxter, T. (2022). Validation of genomic and transcriptomic models of homologous recombination deficiency in a real-world pan-cancer cohort. *BMC Cancer*, *22*(1), 587. https://doi.org/10.1186/s12885-022-09669-z

Li, G.-M. (2008). Mechanisms and functions of DNA mismatch repair. *Cell Research*, *18*(1), 85–98. https://doi.org/10.1038/cr.2007.115

Lieber, M. R. (2010). The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annual Review of Biochemistry*, *79*, 181–211. https://doi.org/10.1146/annurev.biochem.052308.093131

Lindahl, T., & Barnes, D. E. (2000). Repair of endogenous DNA damage. *Cold Spring Harbor Symposia on Quantitative Biology*, *65*, 127–133. https://doi.org/10.1101/sqb.2000.65.127

Livasy, C. A., Karaca, G., Nanda, R., Tretiakova, M. S., Olopade, O. I., Moore, D. T., & Perou, C. M. (2006). Phenotypic evaluation of the basal-like subtype of invasive breast carcinoma. *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc*, *19*(2), 264–271. https://doi.org/10.1038/modpathol.3800528

Lord, C. J., & Ashworth, A. (2017). PARP inhibitors: Synthetic lethality in the clinic. *Science*, *355*(6330), 1152–1158. https://doi.org/10.1126/science.aam7344

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Marquard, A. M., Eklund, A. C., Joshi, T., Krzystanek, M., Favero, F., Wang, Z. C., Richardson, A. L., Silver, D. P., Szallasi, Z., & Birkbak, N. J. (2015). Pan-cancer analysis of genomic scar signatures associated with homologous recombination deficiency suggests novel indications for existing cancer drugs. *Biomarker Research*, *3*, 9. https://doi.org/10.1186/s40364-015-0033-4

Mateo, J., Carreira, S., Sandhu, S., Miranda, S., Mossop, H., Perez-Lopez, R., Nava Rodrigues, D., Robinson, D., Omlin, A., Tunariu, N., Boysen, G., Porta, N., Flohr, P., Gillman, A., Figueiredo, I., Paulding, C., Seed, G., Jain, S., Ralph, C., … de Bono, J. S. (2015). DNA-Repair Defects and Olaparib in Metastatic Prostate Cancer. *The New England Journal of Medicine*, *373*(18), 1697–1708. https://doi.org/10.1056/NEJMoa1506859

Moore, K., Colombo, N., Scambia, G., Kim, B.-G., Oaknin, A., Friedlander, M., Lisyanskaya, A., Floquet, A., Leary, A., Sonke, G. S., Gourley, C., Banerjee, S., Oza, A., González-Martín, A.,

Aghajanian, C., Bradley, W., Mathews, C., Liu, J., Lowe, E. S., … DiSilvestro, P. (2018). Maintenance Olaparib in Patients with Newly Diagnosed Advanced Ovarian Cancer. *The New England Journal of Medicine*, *379*(26), 2495–2505. https://doi.org/10.1056/NEJMoa1810858

Muniandy, P. A., Liu, J., Majumdar, A., Liu, S.-T., & Seidman, M. M. (2010). DNA interstrand crosslink repair in mammalian cells: step by step. *Critical Reviews in Biochemistry and Molecular Biology*, *45*(1), 23–49. https://doi.org/10.3109/10409230903501819

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

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L. B., Martin, S., Wedge, D. C., Van Loo, P., Ju, Y. S., Smid, M., Brinkman, A. B., Morganella, S., Aure, M. R., Lingjærde, O. C., Langerød, A., Ringnér, M., … Stratton, M. R. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, *534*(7605), 47–54. https://doi.org/10.1038/nature17676

Noordermeer, S. M., Adam, S., Setiaputra, D., Barazas, M., Pettitt, S. J., Ling, A. K., Olivieri, M., Álvarez-Quilón, A., Moatti, N., Zimmermann, M., Annunziato, S., Krastev, D. B., Song, F., Brandsma, I., Frankum, J., Brough, R., Sherker, A., Landry, S., Szilard, R. K., … Durocher, D. (2018). The shieldin complex mediates 53BP1-dependent DNA repair. *Nature*, *560*(7716), 117–121. https://doi.org/10.1038/s41586-018-0340-7

O'Connor, M. J. (2015). Targeting the DNA Damage Response in Cancer. *Molecular Cell*, *60*(4), 547–560. https://doi.org/10.1016/j.molcel.2015.10.040

Okoniewski, M. J., & Miller, C. J. (2006). Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, *7*, 276. https://doi.org/10.1186/1471-2105-7-276

Pannunzio, N. R., Li, S., Watanabe, G., & Lieber, M. R. (2014). Non-homologous end joining often uses microhomology: implications for alternative end joining. *DNA Repair*, *17*, 74–80. https://doi.org/10.1016/j.dnarep.2014.02.006

Peng, G., Chun-Jen Lin, C., Mo, W., Dai, H., Park, Y.-Y., Kim, S. M., Peng, Y., Mo, Q., Siwko, S., Hu, R., Lee, J.-S., Hennessy, B., Hanash, S., Mills, G. B., & Lin, S.-Y. (2014). Genome-wide transcriptome profiling of homologous recombination DNA repair. *Nature Communications*, *5*, 3361. https://doi.org/10.1038/ncomms4361

Polak, P., Kim, J., Braunstein, L. Z., Karlic, R., Haradhavala, N. J., Tiao, G., Rosebrock, D., Livitz, D., Kübler, K., Mouw, K. W., Kamburov, A., Maruvka, Y. E., Leshchiner, I., Lander, E. S., Golub, T. R., Zick, A., Orthwein, A., Lawrence, M. S., Batra, R. N., … Getz, G. (2017). A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nature Genetics*, *49*(10), 1476–1486. https://doi.org/10.1038/ng.3934

Popova, T., Manié, E., Rieunier, G., Caux-Moncoutier, V., Tirapo, C., Dubois, T., Delattre, O., Sigal-Zafrani, B., Bollet, M., Longy, M., Houdayer, C., Sastre-Garau, X., Vincent-Salomon, A., Stoppa-Lyonnet, D., & Stern, M.-H. (2012). Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Research*, *72*(21), 5454–5462. https://doi.org/10.1158/0008-5472.CAN-12-1470

Prat, A., Cruz, C., Hoadley, K. A., Díez, O., Perou, C. M., & Balmaña, J. (2014). Molecular features of the basal-like breast cancer subtype based on BRCA1 mutation status. *Breast Cancer Research and Treatment*, *147*(1), 185–191. https://doi.org/10.1007/s10549-014-3056-x

Pusztai, L., Yau, C., Wolf, D. M., Han, H. S., Du, L., Wallace, A. M., String-Reasor, E., Boughey, J. C., Chien, A. J., Elias, A. D., Beckwith, H., Nanda, R., Albain, K. S., Clark, A. S., Kemmer, K., Kalinsky, K., Isaacs, C., Thomas, A., Shatsky, R., … Esserman, L. J. (2021). Durvalumab with olaparib and paclitaxel for high-risk HER2-negative stage II/III breast cancer: Results from the adaptively randomized I-SPY2 trial. *Cancer Cell*, *39*(7), 989–998.e5. https://doi.org/10.1016/j.ccell.2021.05.009

Scully, R., Panday, A., Elango, R., & Willis, N. A. (2019). DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nature Reviews. Molecular Cell Biology*, *20*(11), 698–714. https://doi.org/10.1038/s41580-019-0152-0

Severson, T. M., Wolf, D. M., Yau, C., Peeters, J., Wehkam, D., Schouten, P. C., Chin, S.-F., Majewski, I. J., Michaut, M., Bosma, A., Pereira, B., Bismeijer, T., Wessels, L., Caldas, C., Bernards, R., Simon, I. M., Glas, A. M., Linn, S., & van 't Veer, L. (2017). The BRCA1ness signature is associated significantly with response to PARP inhibitor treatment versus control in the I-SPY 2 randomized neoadjuvant setting. *Breast Cancer Research: BCR*, *19*(1), 99. https://doi.org/10.1186/s13058-017-0861-2

Sfeir, A., & Symington, L. S. (2015). Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? *Trends in Biochemical Sciences*, *40*(11), 701–714. https://doi.org/10.1016/j.tibs.2015.08.006

Shi, Z., Chen, B., Han, X., Gu, W., Liang, S., & Wu, L. (2023). Genomic and molecular landscape of homologous recombination deficiency across multiple cancer types. *Scientific Reports*, *13*(1), 8899. https://doi.org/10.1038/s41598-023-35092-w

Sokol, E. S., Pavlick, D., Khiabanian, H., Frampton, G. M., Ross, J. S., Gregg, J. P., Lara, P. N., Oesterreich, S., Agarwal, N., Necchi, A., Miller, V. A., Alexander, B., Ali, S. M., Ganesan, S., & Chung, J. H. (2020). Pan-Cancer Analysis of BRCA1 and BRCA2 Genomic Alterations and Their Association With Genomic Instability as Measured by Genome-Wide Loss of Heterozygosity. *JCO Precision Oncology*, *4*, 442–465. https://doi.org/10.1200/po.19.00345

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lønning, P. E., Brown, P. O., Børresen-Dale, A.-L., & Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene

expression data sets. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(14), 8418–8423. https://doi.org/10.1073/pnas.0932692100

Steele, C. D., Abbasi, A., Islam, S. M. A., Bowes, A. L., Khandekar, A., Haase, K., Hames-Fathi, S., Ajayi, D., Verfaillie, A., Dhami, P., McLatchie, A., Lechner, M., Light, N., Shlien, A., Malkin, D., Feber, A., Proszek, P., Lesluyes, T., Mertens, F., … Pillay, N. (2022). Signatures of copy number alterations in human cancer. *Nature*, *606*(7916), 984–991. https://doi.org/10.1038/s41586-022-04738-6

Stewart, M. D., Merino Vega, D., Arend, R. C., Baden, J. F., Barbash, O., Beaubier, N., Collins, G., French, T., Ghahramani, N., Hinson, P., Jelinic, P., Marton, M. J., McGregor, K., Parsons, J., Ramamurthy, L., Sausen, M., Sokol, E. S., Stenzinger, A., Stires, H., … Allen, J. (2022). Homologous Recombination Deficiency: Concepts, Definitions, and Assays. *The Oncologist*, *27*(3), 167–174. https://doi.org/10.1093/oncolo/oyab053

Swain, S. (2008). Triple-negative breast cancer: metastatic risk and role of platinum agents. *44th Annual Meeting of the American Society of Clinical Oncology*.

Telli, M. L., Timms, K. M., Reid, J., Hennessy, B., Mills, G. B., Jensen, K. C., Szallasi, Z., Barry, W. T., Winer, E. P., Tung, N. M., Isakoff, S. J., Ryan, P. D., Greene-Colozzi, A., Gutin, A., Sangale, Z., Iliev, D., Neff, C., Abkevich, V., Jones, J. T., … Richardson, A. L. (2016). Homologous Recombination Deficiency (HRD) Score Predicts Response to Platinum-Containing Neoadjuvant Chemotherapy in Patients with Triple-Negative Breast Cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *22*(15), 3764–3773. https://doi.org/10.1158/1078-0432.CCR-15-2477

Tutt, A., Tovey, H., Cheang, M. C. U., Kernaghan, S., Kilburn, L., Gazinska, P., Owen, J., Abraham, J., Barrett, S., Barrett-Lee, P., Brown, R., Chan, S., Dowsett, M., Flanagan, J. M., Fox, L., Grigoriadis, A., Gutin, A., Harper-Wynne, C., Hatton, M. Q., … Bliss, J. M. (2018). Carboplatin in BRCA1/2-mutated and triple-negative breast cancer BRCAness subgroups: the TNT Trial. *Nature Medicine*, *24*(5), 628–637. https://doi.org/10.1038/s41591-018-0009-7

Van Houten, B., Santa-Gonzalez, G. A., & Camargo, M. (2018). DNA repair after oxidative stress: current challenges. *Current Opinion in Toxicology*, *7*, 9–16. https://doi.org/10.1016/j.cotox.2017.10.009

Vollebergh, M. A., Lips, E. H., Nederlof, P. M., Wessels, L. F. A., Wesseling, J., Vd Vijver, M. J., de Vries, E. G. E., van Tinteren, H., Jonkers, J., Hauptmann, M., Rodenhuis, S., & Linn, S. C. (2014). Genomic patterns resembling BRCA1- and BRCA2-mutated breast cancers predict benefit of intensified carboplatin-based chemotherapy. *Breast Cancer Research: BCR*, *16*(3), R47. https://doi.org/10.1186/bcr3655

Wang, D., & Lippard, S. J. (2005). Cellular processing of platinum anticancer drugs. *Nature Reviews. Drug Discovery*, *4*(4), 307–320. https://doi.org/10.1038/nrd1691

Wang, J. C. (2002). Cellular roles of DNA topoisomerases: a molecular perspective. *Nature Reviews. Molecular Cell Biology*, *3*(6), 430–440. https://doi.org/10.1038/nrm831

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Wright, W. D., Shah, S. S., & Heyer, W.-D. (2018). Homologous recombination and the repair of DNA double-strand breaks. *The Journal of Biological Chemistry*, *293*(27), 10524–10535. https://doi.org/10.1074/jbc.TM118.000372

Yu, C., Mannan, A. M., Yvone, G. M., Ross, K. N., Zhang, Y.-L., Marton, M. A., Taylor, B. R., Crenshaw, A., Gould, J. Z., Tamayo, P., Weir, B. A., Tsherniak, A., Wong, B., Garraway, L. A., Shamji, A. F., Palmer, M. A., Foley, M. A., Winckler, W., Schreiber, S. L., … Golub, T. R. (2016). High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. *Nature Biotechnology*, *34*(4), 419–423. https://doi.org/10.1038/nbt.3460

Zhang, C., Xu, C., Gao, X., & Yao, Q. (2022). Platinum-based drugs for cancer therapy and anti-tumor strategies. *Theranostics*, *12*(5), 2115–2132. https://doi.org/10.7150/thno.69424

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS One*, *9*(1), e78644. https://doi.org/10.1371/journal.pone.0078644