Multi-scale characterisation of homologous recombination deficiency in breast cancer

Daniel H. Jacobson[1,2], Shi Pan[1], Jasmin Fisher[2], Maria Secrier[1,*]

[1]UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College London, London, UK

[2]UCL Cancer Institute, Paul O'Gorman Building, University College London, London, UK

* To whom correspondence should be addressed (m.secrier@ucl.ac.uk)

Email addresses:

Daniel H. Jacobson – dan.jacobson.15@ucl.ac.uk

Shi Pan – shi.pan@ucl.ac.uk

Jasmin Fisher – jasmin.fisher@ucl.ac.uk

Maria Secrier – m.secrier@ucl.ac.uk

ABSTRACT

Background

Homologous recombination is a robust, broadly error-free mechanism of double strand break repair, and deficiencies cause reliance on alternative repair processes, leading to sensitivity to PARP inhibitors. Patients displaying homologous recombination deficiency can be identified using 'mutational signatures'. However, these patterns are difficult to reliably infer from exome sequencing. Additionally, as mutational signatures are a historical record of mutagenic processes, this limits their utility in describing the current status of a tumour.

Results

Here, we apply two methods for classifying homologous recombination deficiency in breast cancer to explore the features and heterogeneity associated with this phenotype. Firstly, we apply a likelihood-based method which leverages small insertions and deletions to improve classification of homologous recombination deficiency in exome sequenced breast cancers. We find that whilst BRCA+ homologous recombination deficient samples display strong similarities to those harbouring BRCA defects, they deviate in microenvironmental features such as hypoxic signalling. Secondly, using this classification we develop a 228-gene transcriptional signature which simultaneously characterises homologous recombination deficiency and BRCA1/2-defect status, and is linked with PARP inhibitor response. Finally, we apply this signature to single-cell RNA-sequenced breast cancer cohorts to study homologous recombination deficiency at single cell resolution and demonstrate that these cells present a distinct milieu of interactions with immune cells in their microenvironment compared to their HR proficient counterparts, which could inform checkpoint blockade for HRD tumours.

Conclusions

We apply multi-scale approaches to characterise homologous recombination deficiency in breast cancer through the development of mutational and transcriptional signatures. We show that indels, even at low levels, can improve homologous recombination deficiency classification. Additionally, we demonstrate the heterogeneity of homologous recombination deficiency, especially in relation to BRCA status, and show that indications of this feature can be captured at a single-cell level, enabling further investigations into interactions between DNA repair deficient cells and their tumour microenvironment.

Keywords: homologous recombination deficiency, mutational signatures, PARP inhibition, breast cancer, transcriptional classifier, single cell, tumour microenvironment

BACKGROUND

Maintaining genomic integrity is an essential process for ensuring the sustained survival of cancer cells and is enabled via a complex network of pathways forming the DNA damage response (DDR)(1,2). Homologous recombination (HR) is a robust method of repairing double strand breaks, and HR deficiency (HRD) causes the cell to become dependent on alternative repair processes such as non-homologous end joining (NHEJ) and theta-mediated end joining (TMEJ)(3). These reliances provide therapeutic opportunities to target HRD cells with treatments such as poly ADP-ribose polymerase (PARP) and polymerase θ (Polθ) inhibitors(4–7). PARP inhibitors have already been accepted for clinical use for patients with breast and ovarian cancer harbouring mutations in *BRCA1* and *BRCA2*(8–10). However, a significant proportion of cancer patients appear to display evidence of HRD without harbouring these biomarkers(11–13). Consequently, identifying patients demonstrating HRD, who may therefore benefit from these therapies, has received close attention.

One method for identifying HRD is using patterns of genomic aberrations known as 'mutational signatures'(14). These signatures act as markers of prior mutagenic events and features, and identifying them within the cancer genome can highlight the driving forces behind the development of a given tumour. Signatures of HRD were uncovered as early as initial studies of mutational signatures(15), and HRD has since been linked with signatures of single-base substitutions (SBS), small insertions and deletions (indels), and structural variants(16,17). Specific classifiers have also been created involving the integration of multiple signatures(11) and individual mutational events(12). However, these classifiers work most optimally when applied to whole genome sequenced (WGS) data, and do not perform as well when capturing a reduced representation of the genome e.g. through exome sequencing, which is a common feature of large-scale cancer genomics datasets such as The Cancer Genome Atlas (TCGA).

Alternatives have included measures of chromosomal instability such as copy number signatures, which have been identified in a range of cancers(18–21) as well as the Myriad HRD index score(22,23), which integrates multiple large-scale genomic events to provide a broader measure of HRD(24–26). The Signature Multivariate Analysis (SigMA) computational tool applied a

likelihood approach to classify both exomes and targeted panel sequenced data as SBS3-enriched(27), which has been shown to successfully predict olaparib response in breast and ovarian cancers(28). However, the scope of SigMA has been limited to the SBS3 signature alone as an HRD marker, and the inclusion of HRD-associated indel events offers potential for improved classification.

Whilst mutational signatures of HRD have shown substantial potential for predicting PARP inhibitor sensitivity, one drawback of this method is that, because these signatures result from the accumulation of mutations induced by various carcinogens acting from the very early stages of cancer initiation, they are by definition a feature of the history of a tumour, as opposed to its current state. This is of particular importance when it comes to HRD, as mechanisms of HR revival, such as BRCA1 reversion mutations and loss of 53BP1, lead to the emergence of PARP inhibitor resistance in BRCA1-defective patients(29,30). A signature based on the expression profile of a sample would, therefore, be more adept at describing the most recent state of a tumour. Transcriptional signatures have been applied to characterise various features associated with HRD, including BRCA1 loss in patients(31), HR gene knockdown in cell lines(32), PARP inhibitor sensitivity(33), and chromosomal instability(34). Additionally, the presence of signature SBS3 has been used to develop a gene signature for HRD classification in TNBC(35). This disregards the divergent consequences of BRCA1 and BRCA2 defects due to their different roles in governing HR function: whilst BRCA2 is intrinsically involved in HR via the recruitment of RAD51 to exposed single-stranded DNA, BRCA1 functions upstream of HR and is involved in determining the choice of repair pathway by inhibiting 53BP1, which drives end protection and therefore NHEJ(36,37). A transcriptional signature reflecting this heterogeneity may prevent a skewness towards BRCA1-type HRD, whilst also shedding light on the emergence of HRD in BRCA-positive patients.

In order to perform multi-scale characterisation and explore the heterogeneity of HRD, we present a mutational signature-based classifier of HRD for exome sequenced breast cancers which we then apply to develop a transcriptional signature of HRD and BRCA1/2 deficiency. We demonstrate that leveraging HRD-associated indel events improves HRD classification in downsampled WGS samples and in characterising BRCA-defective samples from the TCGA-BRCA cohort as HRD. Additionally, whilst BRCA+ and BRCA-defective HRD samples are broadly similar regarding standard hallmarks of HRD, BRCA+ samples show deviations in mutational and phenotypic

features such as a comparative decrease in hypoxia. Using matched RNA-sequencing data, we then employ multinomial elastic net logistic regression to develop a 228-gene transcriptional signature which can be used to simultaneously predict BRCA1/2-deficiency and HRD status, and is linked with response to PARP inhibitors in both cell lines and patients from the I-SPY2 trial(38). Finally, we apply the signature to explore HRD at the cellular level from single cell RNA-sequencing data, and demonstrate substantial deviations in patterns of crosstalk with the tumour microenvironment (TME) between HRD and HR-proficient tumour cells. Together, these findings demonstrate the value of multi-scale examination of complex phenotypes like HRD and offer opportunities to improve research into the causes and consequences of such deficiencies in human cancers.

RESULTS

Establishing HRD-associated signature phenotypes in whole genome sequenced breast cancers

Prior to studying HRD in exome sequenced samples, we aimed to develop an initial understanding of the recurring profiles of mutational signatures seen in breast cancers. This was achieved by calculating the contributions of breast cancer-associated SBS and indel signatures in 614 whole genome sequenced breast cancers obtained from the International Cancer Genome Consortium (ICGC), and then clustering these profiles to established 'signature phenotypes'. According to finite mixture modelling, the ICGC cohort could be grouped into 20 clusters, seven of which were assigned as 'HRD' due to the enrichment of SBS3 and the indel signatures ID6 and ID8 (Supplementary Figure 1a-b; Supplementary Table 1).

As expected, *BRCA*-defective samples were strongly enriched within the seven HRD clusters, as were samples labelled as HRD by either HRDetect(11) or CHORD(12), with all BRCA-defective samples appearing in an HRD cluster (Supplementary Figure 1c). 117/120 (97.5%) of samples with HRDetect scores greater than 0.7 appear in an HRD-associated cluster. Considering samples with known BRCA status, 75/135 (55.6%) of samples within the HRD clusters are BRCA-defective, similar to the 74/120 (61.7%) of samples with HRDetect scores greater than 0.7. This is demonstrative of the fact that, even accounting for defects in a range of HR genes such as *RAD51C* and *PALB2* which would increase the precision of these classifiers, the source of HRD for many

samples is unknown, and therefore the specificity for any HRD classifier is not expected to reach 100%.

Interestingly, BRCA1 and BRCA2-defective samples could also be broadly separated. Whilst this has been demonstrated using CHORD(12), we show that it can also be achieved using mutational signatures. BRCA2-defective samples are enriched in clusters characterised by increased contribution of the ID6 signature, referred here as 'BRCA2-type HRD'. The ID6 signature displays deletions at microhomologous regions flanking double strand breaks, indicative of high TMEJ activity(39). Alternatively, BRCA1-defective samples appear in clusters featuring increased ID8 signature contributions, which we call 'BRCA1-type HRD', associated with non-homologous end joining(17). Since BRCA1 is heavily involved in determining how a double strand break will be processed, BRCA1-defective samples will naturally rely on non-homologous end joining to be their primary method of DSB repair. Each type of HRD cluster shows high sensitivity for classifying their respective BRCA defect, with BRCA1 and BRCA2 defects being correctly classified with sensitivities of 68.9% and 70.0% respectively (Supplementary Figure 1d). Whilst CHORD shows greater specificity (73.3% and 93.3% respectively) there is concordance between the BRCA-type HRD clustering and CHORD classification (78.3% for BRCA1-type HRD and 72.2% for BRCA2-type HRD). Thus the indel signatures may not only be useful as HRD-associated signatures, but also shed light on the method employed by a sample for tolerating that type of HRD.

Evaluating HRD in exome sequenced breast cancers

A minimum of 50 mutations are generally believed to be required for reliable mutational signature inference(40). By this criterion, SBS signature analysis is unsuitable in 558/968 whole-exome sequenced samples from the TCGA-BRCA cohort, and indel signature analysis is unsuitable in 958/968 (Supplementary Figure 2). In particular, these samples display a median indel load of three, and a mean of 5.39, with 94 samples harbouring zero indel events. There is an opportunity to overcome such limitations when it comes to identifying HRD in exome sequenced cancers by employing the previously described signature phenotypes in whole cancer genomes: rather than identifying the relevant signatures themselves, the signature profile of a cluster can be predicted instead. To achieve this, we developed a likelihood-based computational method which would enable the assignment to the 20 signature phenotypes without the need to calculate the prevalence of specific signatures (Figure 1a). For each of the 20 clusters, a mean mutational spectrum was

160 calculated, representing a 'ground-truth'/baseline profile against which new sample spectra can be
161 compared (Supplementary Figure 3). Subsequently, each mutation in the profile of a new sample
162 will influence the likelihood of that sample being assigned to each of the signature phenotypes,
163 with the prior probabilities of assignment to each cluster determined by their size in the ICGC-
164 BRCA cohort.

165 To initially demonstrate the capacity of this technique and the accuracy gained by considering indel
166 events, we simulated exome sequenced samples by downsampling whole-genomes from ICGC
167 data to see if our framework would correctly classify the downsampled data according to clusters
168 defined by mutational signatures (see Methods). For these simulations, we based the clustering
169 solely on SBS signatures as a conservative approach that would avoid any bias from indel events
170 which might favour our methodology. We sampled events from each ICGC sample with
171 replacement, constraining them to specific proportions of indel events to demonstrate how their
172 inclusion improved classification. Simulated exomes were set to 25, 50, and 100 mutations. At
173 mutation loads of 50, classification of SBS3 enrichment improved when 5-10% of the simulated
174 mutations were constrained to indel events, indicating that even a small number of indels could
175 improve HRD classification (Fig. 1b). As the simulated mutational load decreased from 100 to 25,
176 a larger proportion of indels was required to enable substantial improvement over using SBS events
177 alone (5% versus 20%, Supplementary Figure 4). The mean and median indel proportions across
178 the TCGA-BRCA cohort are 6.75% and 5.88% respectively, thereby demonstrating that the results
179 of these simulations align with real-world features and that the inclusion of indel events improves
180 HRD classification.

181 Given the demonstrated and known importance of indel events for HRD classification, we sought
182 to understand whether increasing the weights of indels within the likelihood distributions would
183 improve classification. The above analysis was repeated in WGS samples downsampled to 50
184 mutations, and the respective indel proportions within the likelihood distributions were multiplied
185 by factors increasing from 1/5 to 5 (see Methods). We demonstrate that whilst HRD classification
186 is broadly unaffected when the indel likelihood weight is decreased, accuracy decreases
187 significantly once it is increased, thereby showing that overestimating the importance of these indel
188 events will hamper the accuracy of HRD classification (Supplementary Figure 5). Thus, from our
189 analysis we conclude that the optimal HRD classification is attained when indel spectra are equally
190 weighted as the SNV ones.

191 Finally, we asked if the full set of 20 phenotypes seen in WGS data could be captured at lower
192 mutational loads. To this end, we conducted simulations to test the reclassification ability for each
193 phenotype following subsampling (Supplementary Figure 6). This demonstrated that
194 reclassification of specific APOBEC-enrichment clusters was highest at low mutational loads,
195 followed by HRD-enrichment. Additionally, in cases where a sample was misclassified, it was
196 most often assigned to a broader associated cluster. Thus the APOBEC and HRD phenotypes are
197 reliably identified as such even when the mutation load in a tumour is low or not fully captured by
198 the sequencing technique.

199 We next applied the likelihood-based classifier to 968 exome-sequenced breast cancer samples
200 obtained from TCGA (Figure 1c; Supplementary Table 2). The HR gene-defective samples were
201 mainly assigned across the seven HRD-associated clusters. The overall probability of HRD
202 classification was calculated as the sum of probabilities of assignment to the seven HRD-associated
203 clusters, and samples with a classification probability greater than 0.79 were labelled as HRD. This
204 value was selected to maximise the accuracy of the classifier for identifying patients with defects
205 or alterations in HR genes (BRCA1, BRCA2, RAD51C, and PALB2, hereafter termed "HR gene-
206 defective"), as determined using an F-score (Methods; Supplementary Figure 7a). This method was
207 selected to ensure sufficient sensitivity for classifying samples harbouring known HR gene defects,
208 whilst maximising the confidence in HRD classification of HR gene-positive samples. Overall, HR
209 gene-defective samples from TCGA were labelled as HRD with an AUC of 0.91 (Figure 1d,
210 Supplementary Figure 7b), 73.4% sensitivity and 74.8% specificity (Figure 1e).

211 In general, the accuracy of the classifier is greater when considering ER-negative samples
212 compared to ER-positive ones, likely due to the increased enrichment of HR gene-defective
213 samples within ER-negative samples (31.6% versus 6.79%) (Supplementary Figure 7c-d,
214 Supplementary Table 2). It is known that HRD features can vary substantially between cancer
215 types(41), and therefore in some cases redefining specific thresholds based on a given context can
216 improve classification. However, in both cases, the probability threshold of 0.79 performs well for
217 maximising the resulting F-scores.

218 A similar performance was observed when applying the method in an independent exome-
219 sequencing validation dataset of 186 breast cancer patients from the SMC Korean breast cancer
220 cohort(42). Here, 13/20 (65%) of BRCA-defective patients were classified as HRD, increasing to
221 17/20 (85%) when applying a probability cut-off of 0.5 (Supplementary Figure 8).

222 Our method outperformed other signature-based methods such as SigMA(27), the CX3 copy
223 number signature from Drews et al(20), or SBS3-based identification alone in terms of both
224 specificity and sensitivity for classifying HR gene-defective samples as HRD (Figure 1f). It also
225 outperformed the HRD index score, which is based on the levels of loss of heterozygosity, large
226 scale transitions and telomeric allelic imbalances in a sample (Figure 1f). We note that the HRD
227 large-scale genomic aberration features required for the Myriad HRD test are often imperfectly
228 called from exome-sequencing data, and that efforts have been made to optimise the calling of
229 these features(41). Therefore, we have used two different thresholds of HRD classification using
230 this HRD index score based on literature-reported cutoffs(43,44). Whilst our classifier had a recall
231 in TCGA of only 73.4%, generally lower than some alternative methods proposed, it demonstrated
232 a far superior precision of 74.8%, demonstrating an increased stringency and confidence for HRD
233 classification (Supplementary Figure 9).

234 Assignment to HRD clusters occurs more frequently for BRCA2-defective samples (93%)
235 compared with BRCA1-defective (72%) and RAD51-defective (68%) samples (Supplementary
236 Figure 10). However, whilst HRD classification was strong amongst HR gene-defective samples,
237 unlike in the ICGC data, the classifier was unable to assign BRCA-defective samples to BRCA1-
238 or BRCA2-type HRD clusters specifically, with respective sensitivities of 55.0% and 37.0%
239 (Supplementary Figure 10). This is likely due to the scarcity of indel events appearing in exome-
240 sequenced samples.

241 **Hallmarks of HRD**

242 Primary TCGA-BRCA samples labelled as HRD according to the classifier displayed numerous
243 hallmarks associated with DDR deficiencies (Figure 2), including greater levels of large-scale
244 genomic scarring (Fig. 2a), and high levels of the CX3 copy number signature (Fig. 2b), which has
245 been linked with impaired HRD alongside increased replication stress and impaired damage
246 sensing(20), and which performed particularly well for HR gene-defect HRD classification
247 (Supplementary Figure 11). HRD samples displayed increased expression of POLQ (Fig. 2c),
248 indicating a greater reliance on theta-mediated end joining for double strand break repair(6,45,46),
249 as well as increased proliferative capacity (Fig. 2d), which is expected to be associated with
250 HRD(47).

251 Within TCGA, we see that HRD samples are significantly enriched amongst TNBC samples, with
252 38.3% of TNBC samples labelled as HRD (with TNBC constituting 40.9% of all HRD samples)
253 compared with 7.1% amongst the remaining samples (Fig. 2e-f), which was also observed in the
254 SMC breast cancer cohort(42). This aligns with prior information pointing to TNBC samples in
255 TCGA being highly enriched for HR gene defects (43.2% compared with 7.2% amongst the
256 remaining samples). We also classify 55% of the TNBC samples in the SMC cohort as HRD, which
257 is aligned with their findings that 85% of these samples displayed at least some SBS3 signature
258 contribution (Supplementary Figure 8c). Finally, we observed an association between HRD and
259 amplification of MYC (Fig. 2g), which could imply an increase in replication stress(48).

260 HRD and HR-proficient breast cancers in TCGA were shown to display differential enrichment of
261 mutational drivers of tumorigenesis (Fig. 2h). *TP53* mutations were significantly enriched in HRD
262 samples, in agreement with their common co-occurrence with BRCA1/2 defects(49,50). In
263 contrast, the genes *CDH1*, *MAP3K1*, *PIK3CA*, and *GATA3* were more frequently altered in HR-
264 proficient samples. Mutations in *GATA3*, which is involved in normal mammary gland
265 development and has been previously associated with ER-positivity, occur frequently in breast
266 cancer, in particular frameshift indels(16,51,52), and were strongly enriched in a cohort of Nigerian
267 breast cancer patients(53).

268 We applied the dN/dS method(54) to identify signals of positive selection for mutations within the
269 HRD and HR-proficient groups (Fig. 2i). Unsurprisingly, mutations in *TP53* and *PIK3CA* were
270 positively selected within both groups, acting as generic drivers in this cancer. Seven genes were
271 positively selected in the HR-proficient group but not HRD. In contrast, *CASP8* and *F5* were
272 positively selected only in HRD samples. Caspase signalling has previously been hypothesised as
273 a driver of cell death following the induction of cGAS-STING and interferon signalling induced
274 by knock-out of BRCA2, indicating that positive selection for *CASP8* mutation could allude to a
275 method of maintaining cell viability in the context of increased chromosomal instablility(55).

276 BRCA1- and BRCA2-defective samples showed different patterns of positive selection according
277 to dN/dS analysis, with BRCA1-defective and HRD BRCA+ samples showing positive selection
278 only for *TP53*, whilst BRCA2-defective samples only demonstrated this in *PIK3CA* (Fig. 2i). (56)

279 Due to the chromosomal instability associated with HRD, we also sought to highlight associated
280 copy number aberrations. Chromosome arms were more likely to be enriched for losses than gains

281 in HRD samples, aligning our knowledge of loss of heterozygosity as an HRD signature(26,56)
282 (Fig 2j). Only chromosome arm 16p was significantly enriched for both losses in HRD samples,
283 and gains in HR-proficient samples. Notably, the 16p arm carries the *PALB2* gene, which is
284 strongly involved in HR and has been associated with PARPi sensitivity(57–59). Only three arms
285 were significantly enriched for gains in HRD samples, which were 3q, which carries *POLQ*, 10p,
286 which carries the *DCLRE1C* gene, which encodes for the Artemis protein that is essential for NHEJ
287 activity(60), and 8q, which carries the *SPIDR* and *RAD54B* genes, encoding accessory factors for
288 RAD51 activity during homologous recombination(61,62), indicating a potential compensatory
289 mechanism in these patients.

290 Given the variation in mutation selection across HRD samples, we sought to further analyse how
291 BRCA+ HRD samples compare to those harbouring BRCA defects. HRD-BRCA+ samples
292 showed significantly increased levels of HRD hallmarks in comparison to HR-proficient samples
293 (Supplementary Figure 11a). Whilst they showed a slight decrease in CX3 copy number signature
294 contribution compared with BRCA-defective samples, these samples showed no difference in
295 *POLQ* expression or proliferative capacity, indicating that they were displaying a clear HRD
296 phenotype despite their BRCA+ status. Additionally, patients with an HRD classification
297 probability between 0.5 and 0.79 display significantly lower levels of HRD hallmarks compared
298 with those exceeding the threshold, further indicating increased confidence in our HRD
299 classification (Supplementary Figure 11b).

300 Additionally, we checked the somatic mutation status of HRD HR gene-positive samples in
301 comparison with those carrying HR gene defects. In commonly altered cancer genes, there were
302 no discernible differences between BRCA+ and HR gene-defective samples (Supplementary
303 Figure 12a). Of the 27 HRD HR gene-positive samples, 25 (93%) presented a non-silent mutation
304 in a DDR gene, of which 10 (37%) carried mutations in a double strand repair gene, according to
305 a curated list of genes associated with DDR(63). Whilst DDR gene mutations were similarly
306 frequent in BRCA-defective HRD samples (96%), the proportion of double strand repair-mutated
307 samples, excluding BRCA1 and BRCA2, was significantly higher at 61%. Within HRD BRCA+
308 samples, 17/27 (63%) carried a *TP53* mutation (similar to 68% amongst HR gene-defective
309 samples), with the next most commonly mutated DDR genes being *ARID1A* and *MDM4*, a p53
310 inhibitor, although this was only in 2/27 cases each (Supplementary Figure 12b). Interestingly,
311 mutations in *ARID1A* have previously been associated with PARPi sensitivity(64), indicating a

312 potentially rare cause of HRD. The 16p chromosome arm, carrying the *PALB2* gene, was
313 marginally enriched for gains amongst HRD HR gene-positive samples (Supplementary Figure
314 12c). However, following multiple testing correction, no genes were enriched for gains or losses
315 within HRD samples depending on HR gene defects (Supplementary Figure 12d).

316 To gain a broader perspective on the differences driven by BRCA status in HRD samples, we
317 analysed variation in pathway activity using decoupleR(65). Unsurprisingly, estrogen and p53
318 signaling were significantly downregulated in HRD samples (Fig. 2k). We also found that the
319 hypoxia signalling response was substantially upregulated in HRD compared with HR-proficient
320 samples (Fig. 2k). It has previously been demonstrated that BRCA-defective samples from TCGA
321 display increased hypoxia scores compared with BRCA+ samples(66). Here, we found that HRD
322 HR gene-positive samples also show increased hypoxia scores against HR-proficient samples,
323 although this association disappears when analysing ER-negative samples alone (Fig. 2l).
324 Interestingly, whilst these samples have lower hypoxia scores compared to BRCA1-defective
325 samples, this was not observed in comparison to BRCA2-defective samples. A two-way ANOVA
326 revealed that even after accounting for ER-status, both BRCA-defect ($F(2,801) = 11.28$, $p = 1.5e-$
327 05) and HRD status ($F(1,801) = 11.28$, $p = 8.2e-04$) were significantly associated with hypoxia
328 scores (Supplementary Table 3). Severe hypoxia has been shown to lead to PARPi sensitivity in
329 HR-proficient tumours(67), and hypoxia has also been shown to inhibit ER expression in breast
330 cancer cells(68), potentially explaining the similar hypoxia levels across BRCA+ ER-negative
331 breast cancers. However, hypoxia as a BRCA-independent mechanism of HRD requires further
332 analysis and experimental validation that is beyond the scope of this study.

333 Overall, we confirm that the samples classified as HRD via our signature-based method display
334 numerous HRD-associated hallmarks, and demonstrate that HRD and HR-proficient samples show
335 noticeable variation in genomic profiles. Additionally, HRD samples can show variation depending
336 on HR gene-defect status both at mutational as well as signalling activity level, in particular
337 hypoxia, demonstrating physiological deviations which may be reliant on BRCA defects
338 specifically.

339 **Developing a transcriptional signature of BRCA1/2 deficiency**

340 To further explore the functional consequences of HRD, we sought to develop a transcriptional
341 signature reflecting the gene expression profiles of HRD and HR-proficient breast cancers. We

342 aimed to ensure that the signature encompassed the various forms of HRD that could be driven by
343 different factors, such as BRCA1 or BRCA2 defects as well as BRCA-independent mechanisms
344 (HRD-BRCA+). To this end, we trained a multinomial elastic net regression model on the
345 expression profiles from two thirds of the TCGA-BRCA cohort to distinguish between different
346 forms of HRD or HR proficiency, with the remaining TCGA-BRCA samples used for testing (Fig.
347 3a, Methods). A multinomial elastic net approach was chosen due to the ability to remove
348 uninformative features, whilst also tolerating correlated variables. Additionally, since we were
349 seeking a cancer cell-specific phenotype, as opposed to a signal of the tumour microenvironment
350 which can confound bulk sequenced samples, we conducted expression deconvolution using
351 BayesPrism(69) on the training cohort and used the estimated cancer-specific transcriptional
352 profiles for signature development (see Methods). The estimated cell type fractions from
353 BayesPrism significantly correlated with tumour purity estimates(70), indicating reliable cancer
354 cell-specific estimation (Supplementary Figure 13).

355 The transcriptional signature was generated by conducting 1,000 iterations of multinomial elastic
356 net regression with 10-fold cross validation and extracting the genes which were included in all
357 1,000 iterations (see Methods). An HRD 'score', as well as four scores representing HR-
358 proficiency, BRCA1ness, BRCA2ness, and HRD-BRCA-positivity, were then calculated using a
359 centroid-based approach (see Methods). This procedure was conducted with varying regularisation
360 parameters, as well as with and without observation weights to account for imbalanced groups (see
361 Methods). The final signature, containing 228 genes (Fig. 3b, Supplementary Table 4), was selected
362 on account of its optimal capture in single cell data based on the Qian et al. cohort(71)
363 (Supplementary Figure 14).

364 The resulting HRD score was adept at classifying samples labelled as HRD or HR-proficient
365 according to the mutational signature-based classifier (AUC = 0.88) (Fig. 3c). An advantage of the
366 signature was that whilst BRCA1-defective samples showed the highest HRD scores, BRCA2-
367 defective samples also demonstrated HRD scores greater than HR-proficient samples,
368 demonstrating that this signature can capture overall features of heterogeneous HRD (Fig. 3d). For
369 HRD/HR-proficiency classification, the signature outperformed other transcriptional signatures
370 associated with HRD(32), BRCA1ness(31), chromosomal instability (CIN70)(34), and PARP
371 inhibitor sensitivity (PARPi7)(33), as well as gene markers associated with HRD (Fig. 3e).

372 The signature also showed potential at classifying samples depending on their specific BRCA-
373 defect/HRD status, including when applied to BRCA2-defective samples (Fig. 3f; Supplementary
374 Figure 15). The elastic net signature outperformed all others in characterising BRCA1ness,
375 BRCA2ness and HR-proficiency, further demonstrating that, unlike alternative HRD classifiers,
376 using a multinomial approach prevents the resulting signature from skewing away from
377 BRCA2ness, ensuring that the established HRD heterogeneity is captured. Whilst the overall
378 distribution of HRD scores is greater for ER-negative samples in the TCGA testing data regardless
379 of HRD/HR-proficiency status, HRD is also identifiable using the signature after accounting for
380 breast cancer subtype, suggesting that it is not just merely capturing the ER status of the tumours
381 (Supplementary Figure 16).

382 These results were validated in the SMC breast cancer cohort(42). The 228-gene signature
383 outperformed alternative methods in HRD classification (AUC = 0.81) and demonstrated adept
384 capacity for BRCA-defect and HRD BRCA+ classification in comparison with alternative methods
385 (Supplementary Figure 17). It is noted that HRD classification capacity is slightly reduced in the
386 SMC cohort which is suspected to be due to the small number of BRCA1-defective patients in the
387 SMC cohort, whilst BRCA1-defective patients dominated the HRD group in TCGA.

388 According to gene set enrichment analysis, DNA repair processes are dominantly enriched across
389 the signature, as driven by *BRCA1*, *BRCA2*, *TOP3B*, and *FANCI* (Supplementary Figure 18).
390 Intriguingly, the signature was also enriched for genes associated with insulin signalling and
391 glucose transport (*IRS1*, *IRS2*, *CACNA1D*, *SOCS3*, *PRKCZ*), and autophagy and mTOR signalling
392 (*RPTOR*, *RRAGD*, *GABARAP*, *ATP6V1E2*, *ATP6V1C1*).

393 **Key transcriptional contributors to HRD classification**

394 Whilst the 228-gene signature predicts both the HRD and BRCA defect status, we were interested
395 to explore whether a reduced signature could characterise HRD sufficiently. To achieve this, we
396 employed graph attention networks (GATs) that would help us prioritise genes in the signature
397 which have the greatest contribution to distinguishing HRD and HR-proficient phenotypes (Fig.
398 4a, Methods). The model makes use of the original genes from the signature and the degree to
399 which they are correlated in their expression within the TCGA-BRCA cohort to classify patients
400 as HRD and or HR-proficient, while the correlated gradient information decides which gene sub-
401 modules might drive the classification. Briefly, a patient-specific weighted gene co-expression

402 graph is extracted using weighted correlation network analysis (WGCNA)(72) and these graphs
403 are input into the GAT, which is then trained to distinguish HRD from HR-proficient samples and
404 then selects parts of the graph based on its gradients. The resulting graph neural network showed
405 high accuracy for HRD classification (AUC = 0.90). By ranking the genes using a gradient-based
406 approach for their performance in classifying HRD or HR-proficiency, we were then able to
407 highlight the key genes driving correct model prediction.

408 The analysis highlighted 26 genes which were sufficiently important for classifying the HRD and
409 HR-proficiency groups, of which 15 were predictive of HRD and 11 were predictive of HR-
410 proficiency (Fig. 4b-c, Supplementary Table 5). A number of these genes have been associated
411 with DDR, including *USP13*, a regulator of replication stress involved in ATR activation via
412 TopBP1 deubiquitination(73,74), *POLG*, a key regulator of mitochondrial DNA replication and
413 repair(75,76), and *IP6K2*, a stabiliser of DNA-PKcs and ATM leading to p53
414 phosphorylation(77,78). Additionally, the reduced signature contains two genes encoding zinc
415 finger proteins (*ZNF718* and *ZNF583*) which are involved in both HR and NHEJ(79,80). On their
416 own, these 26 genes provide an adept reduced gene signature for capturing HRD, with maintained
417 capacity for HRD classification across BRCA2-defective samples (Supplementary Figure 19).

418 **Testing the signature against sensitivity to PARP inhibitors**

419 To determine the therapeutic relevance of the transcriptional signature of HRD, we next applied
420 the signature to breast cancer cell lines to predict PARP inhibitor sensitivity. The signature was
421 applied to 68 breast cancer cell lines from the Cancer Cell Line Encyclopaedia (CCLE)(81), and
422 matched to PRISM drug sensitivity data from 26 of these cell lines (Supplementary Table 6).

423 Cell lines with increased transcriptional scores for HRD showed increased sensitivity to four
424 different PARP inhibitors, as represented by lower PRISM scores (Fig. 5a). We note that the
425 correlation is marginal, especially regarding olaparib and talazoparib. However, the 228-gene
426 signature still shows a stronger correlation with PARPi sensitivity in these cell lines than any
427 alternative signatures (Supplementary Figure 20). This is likely a result of the signature being
428 developed from bulk-sequenced primary tumour samples, whilst the cell lines will be lacking a
429 microenvironmental component. Whilst we have attempted to account for microenvironmental
430 signals using expression deconvolution, these extracted cancer cell-specific signals will still exist

431 in a broader environmental context which the cell lines will be lacking, hence a resulting
432 transcriptional HRD signal will likely vary significantly.

433 The HRD score also predicted responses to combined PARP and checkpoint inhibition in breast
434 cancer patients. The signature was applied to 105 HER2-negative, Stage II/III breast cancer patients
435 from the I-SPY2 trial(38). In this trial, 71 patients were treated with a combination of olaparib and
436 the PD-L1 inhibitor durvalumab, as well as the neoadjuvant chemotherapy taxol, whilst 34 control
437 patients were treated with taxol alone. Amongst the patients in the treatment arm, 29 showed
438 pathologic complete response (pCR), and these patients showed significantly increased HRD
439 scores compared to those who did not display pCR (Fig. 5b). Our HRD score was significantly
440 correlated with their own PARPi7 signature score from the trial but showed a better separation
441 between responders and non-responders (Fig. 5c, Supplementary Figure 21). Overall, this
442 demonstrates that, in capturing a general transcriptional phenotype of HRD, the 228-gene HRD
443 signature is also linked with PARP inhibitor sensitivity, especially in patient samples.

444 **Application of the HRD signature to single cells**

445 When considering HRD in the context of treatment or identification, it is often forgotten that whilst
446 HRD can manifest and cause effects at the level of the whole tumour, it is still an intrinsically
447 cellular phenotype. Since the transcriptional signature was developed with the intention of
448 characterising HRD in terms of tumour-specific signalling, as opposed to that of the
449 microenvironment, this suggested that the signature could be applied to scRNA-seq data, which
450 would provide insight into the distribution of HRD across cells and generate further questions about
451 the varying roles of HRD and HR-proficient tumour cells within the context of the tumour
452 microenvironment.

453 To investigate whether the signature may be applicable to single cell data, we first applied it to 11
454 breast cancer samples with matched bulk and single cell-sequencing data from Chung et al(82).
455 We show a good correlation between the bulk HRD scores and the mean HRD scores across the
456 individual tumour cells within each sample (Fig. 6a), providing an indication that the HRD signal
457 captured in bulk sequencing data reflects, on average, the levels seen in single cells.

458 Following this, we applied the signature to a cohort of 14 single cell-sequenced breast cancers
459 containing over 44,000 cells from Qian et al(71). These 14 samples included one BRCA1-defective

460 sample (sc5rJUQ033), which we assumed to be HRD, and one Stage II Luminal A sample
461 (sc5rJUQ064) which we assumed to be HR-proficient, given the characterisation of this type of
462 breast cancer as slow-proliferating(83,84). These two samples display substantially different
463 distributions of HRD scores across the cancer cells in these respective samples, with the Stage II
464 Luminal A sample displaying a distinctively more HR-proficient distribution (Fig. 6b), further
465 demonstrating that a transcriptional signature of HRD can potentially be captured at single-cell
466 resolution.

467 Across the Qian et al. cohort and 31 treatment-naïve samples obtained from the Bassez et al.
468 cohort(85), on average 10.8 and 11.2 genes from the signature were expressed per cancer cell,
469 respectively (Supplementary Figure 22a-b). The mean number of genes in the signature expressed
470 per cell across each samples varied between 6.74 and 17.9 in the Qian et al. cohort (Supplementary
471 Figure 22c) and 6.40 and 23.3 in the Bassez et al. cohort (Supplementary Figure 22d), indicating
472 sufficient capture of the transcriptional signature across the single-cell cohorts.

473 The cancer cells from these 14 samples displayed intra-sample heterogeneity of HRD scores (Fig.
474 6c; Supplementary Figure 23a) that matched the clustering by breast cancer subtype (Fig. 6d).
475 Generally, the triple negative and BRCA1-defective samples presented a greater proportion of
476 HRD cells, defined as displaying a transcriptional score greater than zero (Fig. 6e).

477 Similarly to Qian et al(71), the tumour cells obtained from Bassez et al.(85) displayed a distinctive
478 gradient of HRD scores (Fig. 6f; Supplementary Figure 23b) in accordance with the clustering by
479 breast cancer subtype (Fig. 6g), and TNBC samples displayed greater proportions of HRD cells in
480 comparison with receptor-positive samples(Fig. 6h). Moreover, we found that, whilst the HRD
481 scores from the TME were consistent across samples and tended to centre closely around zero,
482 these scores varied far more broadly across samples within the cancer cells across both cohorts,
483 indicating that the transcriptional signal of HRD is likely arising strongly from the tumour cells
484 (Supplementary Figure 23c-d).

485 This further demonstrates the potential provided by this transcriptional signature of HRD to capture
486 this phenotype at single-cell resolution, as well as demonstrating the heterogeneity of HRD levels
487 across individual samples.

488

489 **Exploration of the HRD tumour microenvironment at single cell level**

490 To further explore the activities of HRD and HR-proficient cells within the tumour
491 microenvironment, we applied CellphoneDB(86), an extensive database of ligand-receptor
492 interactions, to observe whether the interactions established between tumour cells and the
493 surrounding immune and stromal cells varied based on HR capacity utilising both the Qian et al
494 and Bassez et al cohorts(71). In both cohorts, cell-cell interactivity profiles were dominated by
495 fibroblast, endothelial, myeloid, and dendritic cells, with cancer cells demonstrating fewer
496 interactions with the TME (Fig. 7a-b). However, across all cell types, HRD cancer cells
497 consistently displayed fewer significant interactions with the TME, in particular as the target of
498 these interactions, than HR-proficient cells (Fig. 7c-d).

499 While some common interactions were observed, multiple ligand-receptor pairs uniquely mediated
500 the interactions of HR-proficient cells with T cells across both cohorts (Fig. 7e). In particular, these
501 included interactions involving TNF, TGFb, and prostaglandin E2 signalling across both cohorts
502 and from different cell types (Supplementary Figure 24). These results align with a previous
503 hypothesis suggesting downregulation of TNF signalling as a mechanism of cell survival following
504 BRCA2 deficiency on account of the resulting decrease in caspase-induced apoptosis(55). Unique
505 TME-HRD cell interaction pairs often involved *LGALS9* signalling, a feature which also re-
506 occurred across cell types and cohorts (Supplementary Figure 24). *LGALS9* has previously been
507 associated with antimetastatic potential in breast cancer(87), potentially indicating a compensation
508 for increased chromosomal instabilility which may drive metastasis(88). Additionally, expression
509 of the *LGALS9* gene product, Gal-9, is increased following taxane treatment in TNBC, due to
510 nuclear activation of NF-kB, which is also upregulated in HRD breast cancers(89,90).

511 This analysis unveils the specific strategies HRD and HR-proficient cells employ in their crosstalk
512 with the TME, highlighting differences in the type and variety of molecular components involved.
513 Generally, the HRD cells appear less "promiscuous" while HR-proficient cells display a wider
514 array of interaction strategies. However, these findings do not inform us on whether HRD cells
515 respond less or more frequently to other cells in their environment, as the individual cell-cell
516 interactions cannot be inferred using this method, and should not be interpreted as such.

517 Overall, our results highlight a complex pattern of interactions between cancer and non-cancer cells
518 which may be mediated by the DNA damage response and could, in the long term, inform treatment
519 strategies that jointly target HRD tumours and their microenvironments.

520 **DISCUSSION**

521 As the clinical utility and capabilities of personalised treatments increase, it is necessary to ensure
522 that the features predicting positive treatment response can be identified reliably. In this study, we
523 present multi-scale approaches to characterising HRD which can be applied in a variety of contexts.
524 We developed a method for high-confidence HRD identification in exome sequenced breast
525 cancers which incorporates indel events that indicate both the presence of HRD and alternative
526 DSB repair mechanisms employed in the event of HRD. We demonstrate that even small amounts
527 of indels like the ones expected to be seen in WES data improve HRD classification. Furthermore,
528 the HRD group defined by our genomic signatures displays the characteristic features expected of
529 such cancers, including *MYC* amplification and elevated *POLQ* expression. Applying this classifier
530 to the TCGA-BRCA cohort, we then developed a 228-gene transcriptional signature that
531 characterises the heterogeneity of HRD, whilst also correlating with PARP inhibitor sensitivity and
532 displaying the capacity to define HRD at single cell resolution.

533 In creating the mutational signature-based classifier, we only considered SBS and indel signatures
534 due to our focus on the HRD phenotype and data availability in TCGA. However, the
535 generalisability of the method is worth highlighting. Copy number signatures could be effectively
536 integrated into this method to improve HRD classification, especially if copy number profiles can
537 be divided into specific features and contexts as demonstrated in previous studies(20,21).
538 Furthermore, our method also defines subgroups enriched for alternative mutational processes,
539 including APOBEC cytidine deamination and mismatch repair deficiency. With regards to this,
540 double base substitutions (DBSs) can also be included within the method and may be of use for
541 improved classification of processes with associated DBS signatures, such as mismatch repair
542 deficiencies and tobacco-associated mutagenesis.

543 We applied a probability threshold of 0.79 for HRD classification using the mutation-based
544 classifier, which was determined through optimising the resulting F-score for identifying patients
545 with HR gene defects. Whilst using this threshold leads to high-confidence classification of HRD
546 in samples which do not harbour HR gene defects (Supplementary Figure 11b), this does lead to a

547 minor but notable decrease in sensitivity. Additionally, we note that these groups may not be
548 distributed identically in the TCGA cohort, as demonstrated by the differences in rates of BRCA
549 defects (13.6% in ICGC compared to 8.99% in TCGA), suggesting that the prior distributions,
550 whilst based on real data, might not be wholly representative. It is non-trivial whether to emphasize
551 sensitivity or specificity when generating an HRD classifier, in which it is known that the
552 specificity should not be 100%. This is because a key difficulty in developing an HRD classifier is
553 the lack of ground truth beyond HR gene defects and, as was used for SigMA and the simulation
554 analysis conducted in this study, SBS3 signature contribution in WGS data. Therefore, we used a
555 balanced F-score which considers both with equal importance. However, we note that different
556 probability thresholds can be applied, as was utilised for SigMA(27).

557 Mutational signatures have also shown promise for HRD classification in targeted panel
558 sequencing, when presented as a likelihood-based approach as was done through SigMA(27). Due
559 to the availability of gene panels, this development presents invaluable clinical relevance. Whilst
560 we also employ a likelihood-based approach, owing to the substantially decreased indel loads
561 identified through targeted panel sequencing it is unlikely that this method would significantly
562 contribute to improved HRD classification, and therefore do not recommend its application to gene
563 panels. Furthermore, it should be noted that our exome classifier's specific clinical utility is limited.
564 Whole genome sequencing will likely become increasingly available for mutational signature-
565 based diagnostics such as HRDetect, and panel sequencing is already widely applied for identifying
566 gene defects. However, the primary benefit of an exome-based classifier is its application to large-
567 scale genomics resources such as TCGA, enabling further profiling of HRD from a broad range of
568 omics perspectives and new hypotheses generation which these resources enable.

569 Whilst it is not surprising that our 228-gene transcriptional signature outperforms alternative
570 methods given that it was trained using labels determined by our own mutational HRD classifier,
571 this transcriptional signature also simultaneously captures BRCA1- and BRCA2-specific
572 deficiency phenotypes, highlighting the distinct consequences of loss of function of these two
573 genes. It is worth noting that we additionally capture a group of tumours that display transcriptional
574 profiles closer to those of classically HR-deficient BRCA mutated samples but lacking any BRCA
575 defects. These tumours might be experiencing some level of HRD-like state due to more complex
576 changes across the HR and linked pathways, some of which may be epigenetic or of other nature.

577 Further analyses are needed to shed light into the aetiology of these cancers, and it is likely they
578 are a rather heterogeneous group.

579 In terms of its relevance in a therapeutic context, we found that our HRD transcriptional signature
580 was more strongly associated with PARP inhibitor sensitivity in patients than in cell lines. Given
581 that the signature was developed using breast cancer patient samples, this was likely to have been
582 the case. While we have attempted to ensure that we are capturing a tumour-cell intrinsic HRD
583 signature by correcting for microenvironmental signals in bulk data, some tumour intrinsic
584 regulation might still be partly environmentally triggered, and this component would not be
585 captured in cell lines which lack this microenvironment. Additionally, due to a broad variety of
586 factors including genetic instability and growth conditions, drug responses across cell lines may be
587 hugely variable(91), which may partially explain the decrease in signature performance when
588 applied to cell lines.

589 Future developments are likely to focus significantly on PARP inhibitor resistance. Archetypal
590 mechanisms of PARP inhibitor resistance are becoming well established, such as BRCA1 reversion
591 cases, 53BP1 loss following BRCA1 loss, and PARG loss following BRCA2 mutations(92), and
592 features such as reprogramming of cell survival pathways and an increasingly mesenchymal
593 phenotype have been associated with gradual PARP inhibitor resistance(93). Currently there are
594 no available datasets demonstrating the effect of HR resurgence on tumour heterogeneity, however
595 these will provide an invaluable resource for studying HRD moving forward.

596 Additionally, recent developments in inferring mutagenic processes in single cells have shed light
597 on the driving forces behind the evolution of tumour heterogeneity in TNBC and high grade serous
598 ovarian carcinoma (94). In breast cancer, HRD tumours are generally believed to be more
599 immunogenic due to their potential to generate increased mutational loads and neoantigen
600 signalling, which can be exploited for checkpoint inhibition(43,95,96) BRCA defects have also
601 been associated with increased immunosurveillance in high grade serous ovarian cancer, and
602 CellphoneDB was recently applied to highlight a malignant cell population associated with poor
603 prognosis in ovarian cancer, which was associated with immune cell interactions and displayed
604 generally low levels of chromosomal instability(97,98). Additionally, CellphoneDB was used to
605 explore human breast cancer immune microenvironments, and specifically highlighted a large
606 number of unique interactions between fibroblasts and endothelial cells, as well as smaller levels
607 of interaction by T-cells and B-cells, as we have also observed, which was attributed to fewer genes

608 being expressed in these cell types(99). While the TME has been fairly extensively explored in
609 breast cancer bulk datasets and more recently in single cells, our understanding of how the tumour-
610 TME crosstalk is established in the context of HRD or HR proficiency at single cell resolution is
611 much more limited. We show that our transcriptional signature may be employed to highlight
612 patterns of HRD in single cells, and this paves the way for further explorations into the way that
613 DNA repair deficiencies may influence their microenvironments, and vice versa. Thus, we believe
614 our analysis can pave the way to more detailed interaction studies by highlighting specific strategies
615 of tumour-T cell coupling in HRD cells which may have relevance to immunotherapy
616 effectiveness. The association between HRD and various facets of the tumour microenvironment
617 is already being mined for therapeutic potential through investigation of joint PARP and immune
618 checkpoint inhibition(38,100,101), and chromosomal instability has been shown to elicit an
619 inflammatory response offering further targets for combination treatment(55,102). The capacity to
620 investigate differential cell-cell interactivity of HRD and HR-proficient cells may allow for insight
621 into further mechanisms which may be exploited.

622 **CONCLUSIONS**

623 We have demonstrated that HRD classification in exome sequenced breast cancers can be improved
624 by leveraging the presence of HRD-associated indel events, and have shown that mutational and
625 phenotypic profiles of HRD persist regardless of the presence of HR gene defects. These
626 classifications have been used to develop a transcriptional signature which is associated with
627 sensitivity to PARP inhibitors and can be applied to characterise HRD in single-cell RNA-
628 sequencing data. In examining the TME crosstalk at single cell resolution, we demonstrate
629 substantial variation in cell-cell interactivity patterns dictated by the HRD or HR proficiency status
630 of the tumour cells, suggesting distinct pathways mediating immune recognition and/or escape.
631 These findings pave the way to further investigation of the heterogeneity of HRD and HR-
632 proficiency both at patient and individual cell level, as well as therapeutic implications.

633

634

635

636 **MATERIALS AND METHODS**

637 **Data sources**

638 Whole-genome sequencing data from the ICGC-BRCA cohort was obtained from DCC data release
639 28(103). Data from two projects was included: BRCA-UK (n = 45) and BRCA-EU (n = 569),
640 resulting in a total of 614 samples. Exome sequencing data from TCGA was obtained from the
641 GDC Data Portal using the TCGAbiolinks R package(104). Somatic mutation data collated using
642 the Mutect2 pipeline was obtained using the GDCquery() function. Annotation of BRCA-defective
643 samples in TCGA was taken from Valieris et al.(105). Samples were considered BRCA1/2-,
644 RAD51C-, or PALB2-defective if they were assigned either 'Bi-allelic inactivation' or 'Epigenetic
645 silencing'. HRD index scores as calculated by Myriad for TCGA samples were obtained directly
646 from Marquard et al.(22). CX3 scores were obtained from Drews et al.(20).

647 Exome sequencing and transcriptional profiling from the SMC cohort of Korean breast cancer
648 patients , as well as chromosome arm-level copy number alterations for the TCGA-BRCA cohort,
649 were downloaded from cBioPortal(106).

650 FPKM-normalised gene expression data for TCGA was obtained from the GDC Data Portal. The
651 proliferation/cell cycle arrest capacity of the tumours was calculated from RNA-seq profiles using
652 the quiescence (Q) score defined in Wiecek et al. (107), and was defined as 1-Q, with positive
653 scores indicating a higher proliferative capacity.

654 Cell line expression data and PARP inhibitor sensitivity profiles were obtained from the Cancer
655 Cell Line Encyclopaedia (CCLE)(81). Transcriptional profiling and clinical data of patients from
656 the I-SPY2 trial were collected from the Gene Expression Omnibus (GEO) database, with
657 accession number GSE173839.

658 Single-cell RNA-seq data from Chung et al.(108) was collected from GEO database through the
659 accession numbers GSE75688. Treatment-naïve single-cell RNA-seq data from Qian et al.(71) and
660 Bassez et al.(85) were downloaded directly from the Lambrechts laboratory website
661 https://lambrechtslab.sites.vib.be/en/team. Single-cell RNA-seq profiling from Qian et al. and
662 Bassez et al. were processed using the Seurat R package(109), to extract only cancer cells with
663 between 200 and 6,000 unique feature counts and mitochondrial content less than 15%, the
664 expression profiles of which were then log-normalised.

665 **Generation of HRD classifier for exome sequenced breast cancer samples**

666 *Mutational signature contributions in WGS breast cancer samples:*

667 Mutational signature analysis of 614 WGS breast cancer samples obtained from the International
668 Cancer Genome Consortium (ICGC) project was conducted using the deconstructSigs R
669 package(40). SBS and indel signatures from the COSMIC v3.3 database were included if they
670 appeared in >1% breast cancer samples according to the Pan Cancer Analysis of Whole Genomes
671 (PCAWG) project(17). SBS and indel contributions were calculated separately, and the results
672 were combined for subsequent clustering analysis.

673 *Formation of HRD-specific mutational spectra:*

674 The 614 ICGC breast cancer samples were clustered according to their SBS and indel signature
675 contributions, which were calculated separately using the deconstructSigs R package(40). Since
676 we intended to use WGS samples to calculate estimated signature contributions within the exome
677 regions, we normalised the mutation profiles to account for the frequency of each mutation type
678 occurring in the exome relative to the whole genome. For SBS profiles, this was done by setting
679 the option tri.counts.method = 'genome2exome' in deconstructSigs. For indel profiles, we used the
680 ICGC breast cancer cohort to count the total frequency of each of the 83 indel mutation types, and
681 estimated the frequency of each within the exome by excluding mutations appearing within
682 intergenic regions. The indel profiles of each sample were then multiplied by the ratio of the
683 frequency of each mutation type within the exome to the whole genome. Additionally, mutation
684 spectra in exomes may also differ from the rest of the genome on account of transcription-coupled
685 damage and repair, which cannot be accounted for in WGS samples even after factoring for varying
686 triplet frequencies.

687 Model-based clustering was conducted using the mclust R package using finite mixture
688 modelling(110). The final classification and optimal number of clusters was selected according to
689 the Bayesian Information Criterion value. Whilst 22 clusters were initially identified via this
690 method, two of these clusters consisted of only one sample each, neither of which displayed
691 discernible features, therefore these clusters were discarded. The 20 remaining clusters were named
692 according to the most prevalent features, with the seven clusters enriched for SBS3 labelled as
693 'HRD'.

694 The mutational spectrum for each cluster was determined by collectively calculating the mean
695 distribution of the 96 SBS and 83 indel mutation types. The result is 20 representative mutation
696 distributions consisting of mutation events, each summing to one (Supplementary Figure 3).

697 *Application of mutational spectra for HRD classification in TCGA:*

698 The frequencies of the 179 mutation events across 986 exome-sequenced breast cancer samples
699 obtained from TCGA was determined using the sigminer R package(19). The probability of a
700 sample being assigned to a specific cluster given the set of aberrations displayed follows Bayes'
701 theorem as follows:

702 $$P(\, cluster_i \mid S\,) = \frac{P(\, S \mid cluster_i\,) * \; P(\, cluster_i\,)}{P(\, S\,)}$$

703 where S represents the n mutations forming the mutational profile of the sample, $P(\, cluster_i\,)$ is
704 the prior probability of assignment to cluster i, with $i \in [1, 19]$, as estimated from clustering of
705 WGS ICGC samples, and $P(\, S \mid cluster_i\,)$ is the likelihood of S occurring in a sample from cluster
706 i, calculated as:

707 $$P(\, S \mid cluster_i\,) = \prod_{j=1}^{n} P\big(\, s_j \mid cluster_i\,)$$

708 where s_j is the *j*th mutation, $P\big(\, s_j \mid cluster_i\,)$ is the probability of s_j within the mean mutational
709 spectrum of $cluster_i$, and the normalising constant $P(\, S\,)$ is the sum of likelihoods multiplied by
710 prior probabilities for all 19 clusters:

711 $$P(\, S\,) = \sum_{k=1}^{19} P(\, S \mid cluster_k\,) * P(\, cluster_k\,)$$

712 The overall probability of a sample being HRD was calculated as the sum of the probabilities of a
713 sample appearing across the seven HRD clusters, and samples with a probability of greater than
714 0.79 were deemed HR-deficient. For BRCA defect-specific HRD classification, samples are
715 assigned to the specific cluster to which they have the greatest probability of assignment.

716

717 **Evaluation of the HRD classifier**

718 The success of the classifier was determined by calculating its ability to identify patients with
719 BRCA1/2 defects. The characterisation of over 900 TCGA samples for either bi-allelic inactivation
720 or epigenetic silencing of BRCA1 or BRCA2 by Valieris et al.(105) was used as the truth label/gold
721 standard annotation. The F-scores for each HRD classification method were calculated as:

722 $$F = \frac{TP}{TP + \frac{1}{2} * (FP + FN)}$$

723 where TP is the number of true positives, FP is the number of false positives, and FN is the number
724 of false negatives. Whilst we expected model sensitivity to be close to 100% (as we assumed that
725 the majority of BRCA-defective samples would be HRD) and did not expect these results for model
726 specificity (due to the known presence of BRCA-positive HRD samples), we aimed to ensure that
727 overclassification of HR-proficient samples as HRD was as limited as possible, and so did not
728 apply a weighted F-score. HRD classification using the HRD index score was done using
729 thresholds of 42 and 63 due to their application in the literature(23,44).

730 **Simulation Analysis**

731 We tested our method's HRD classification performance in a low mutation count context using
732 simulated data where we varied the fraction of indels included in a sample. This also allowed us to
733 understand what impact indels generally have for HRD classification. Categories for simulation
734 analysis were generated using hierarchical clustering applied to the SBS signature contributions
735 calculated for the ICGC-BRCA cohort. Samples were downsampled to mutational burdens of 25,
736 50, and 100, with the additional constraint that the proportion of indels in the simulated data were
737 set from 0 to 0.5, in steps of 0.05. Each combination of sample sizes and indel proportions was
738 iterated 100 times, and differences in the resulting AUC values for identifying SBS3-enriched
739 samples using the classifier at each iteration were analysed using Wilcoxon rank-sum testing.

740 To analyse the effect of altering the indel contributions within the likelihood distributions, we
741 repeated the analysis except instead of constraining the indel proportions, we multiplied the indel
742 contribution to each likelihood by increasing factors: 1/5, ¼, 1/3, ½, 1, 2, 3, 4, 5.

743 Simulation analysis was repeated using all 20 clusters by downsampling each sample to sizes of
744 25, 50, and 100, with no constraint on indel proportions, calculating the AUC for correct cluster
745 assignment using the classifier, and conducting 100 iterations of this process.

746 **Mutation enrichment analysis**

747 Mutation enrichment analysis was conducted on 738 genes which have been causally implicated
748 in cancer according to COSMIC as of July 3rd 2023 (111). Only genes which were mutated in more
749 than 5% of samples were included in the analysis, and their enrichment in the HRD/HR-proficient
750 groups was calculated using a Fisher's exact test. Mutated genes under positive selection were
751 identified by dN/dS analysis, which was conducted using the dNdScv R package(54) with default
752 parameters. The analysis was run independently for six groups: all HR-proficient, all HRD,
753 BRCA1-defective, BRCA2-defective, RAD51C-defective, and HRD BRCA+.

754 **Chromosome arm-level enrichment analysis**

755 For all chromosome arms, enrichment of CNAs in HRD breast cancer samples compared to HR-
756 proficient samples were calculated using Fisher's exact tests independently testing gains against
757 non-gains (normal or loss), and losses against non-losses (normal or gain).

758 **Pathway Enrichment Analysis**

759 Differential activity of 14 signalling pathways was analysed using decoupleR(112), which was
760 applied to RNA-seq counts from the TCGA-BRCA cohort. Lowly-expressed genes were removed
761 and the remaining data was VSN-normalised. Differential expression analysis was conducted using
762 the DESeq2 R package(113), and the results of this analysis were fed into the decoupleR R package
763 to estimate pathway activity. Cancer type-specific hypoxia scores, using the Buffa signature(114),
764 were obtained from Bhandari et al.(115).

765

766

767 **Generation of a transcriptional signature of HRD**

768 *Data preprocessing*

769 Samples from the TCGA-BRCA cohort were selected only if they included exome-sequencing data
770 (and therefore had been assigned as HRD/HR-proficient), and a BRCA defect label obtained from
771 Valieris et al.(105). To prevent confounding, samples harbouring defects in RAD51C or PALB2
772 were excluded. This resulted in 857 samples, of which two-thirds (n=572) were assigned as training
773 Training and testing sets were defined using the createDataPartition() function from the caret R
774 package(116) to ensure equal proportions of each HRD/BRCA-defect group within the two sets.

775 *Expression deconvolution*

776 Expression deconvolution was conducted using the BayesPrism R package, which estimates cell
777 type-specific bulk expression profiles from a single-cell RNA-seq reference dataset. In this case,
778 the Qian et al.(71) dataset was used as a reference dataset. Genes from selected groups, including
779 mitochondrial, ribosomal, and chromosome X and Y genes, were excluded. Following this, protein
780 coding genes only were included. To ensure that the resulting signature would be suitably
781 applicable to single-cell RNA-seq data, genes that were expressed in less than 2% of the cancer
782 cells in the Qian et al. dataset were excluded from further analysis, resulting in 9,853 genes for
783 downstream analysis.

784 *Development of the multinomial transcriptional signature*

785 The transcriptional signature was generated by multinomial elastic net regularised logistic
786 regression using the glmnet R package(117). We performed 1,000 iterations of 10-fold cross
787 validation using the cv.glmnet() function with type.multinomial = 'grouped'. Initially, four
788 signatures were created, setting alpha = 0.25 or 0.5, and applying or excluding weightings to
789 account for group imbalances. We collated the coefficients provided for all features for each
790 iteration in a model generated using λ = lambda.min being the value of λ which gives the lowest
791 mean cross-validated error. The non-weighted elastic net model, with alpha = 0.25, was selected
792 due to its presence within the Qian et al. cohort (Supplementary Figure 14). The final signature
793 was formed of 228 genes which were assigned non-zero coefficients across all 1,000 iterations.

794 Similarly to Severson et al.(31), the signature was calculated using a nearest centroid method. To
795 create the BRCA1/2-deficiency signature, the TCGA training data was split into its four categories,

796 and a template was created for each group by taking the mean expression of each of the 228 genes
797 across the samples in that category. For new samples, four 'scores' were then created by calculating
798 the Pearson's correlation coefficient between the expression profile of the new sample and each of
799 the four templates.

800 The same procedure was also applied to generate an HRD signature, except only two templates
801 were created relating to 'HRD' and 'HR-proficient'. For a new sample, the Pearson correlation
802 coefficients against the two templates were calculated, and then the correlation with the 'HR-
803 proficient' template was subtracted from the correlation with the 'HRD' template to generate an
804 overall HRD score.

805 This transcriptional signature was compared against four published signatures: a 230-gene HRD
806 signature developed by Peng et al.(32), a 77-gene BRCA1ness signature developed by Severson et
807 al.(31), a 70-gene signature of chromosomal instability (CIN70)(34), and a 7-gene signature of
808 PARPi sensitivity (PARPi7)(33). Application of these signatures for comparison was also
809 conducted using the centroid method described above. In the event of a gene in the signature not
810 appearing in a dataset, the gene was removed from the signature and did not contribute to the
811 correlation calculation.

812 *Gene Set Enrichment Analysis*

813 Gene set enrichment analysis was conducted using the pathfindR R package(118) and enrichR R
814 package(119). For the pathfindR analysis, to provide relevant significance values, an ANOVA was
815 conducted for each gene against the four BRCA-defect groups. The KEGG and Gene Ontology
816 Biological Process gene sets were used, and default inputs were used for the remaining arguments.

817 **Importance Analysis using a Graph Neural Network Approach**

818 A Graph Attention Network (GAT) was used to map gene co-expression graphs into an embedding
819 space and to analyse its classification output in order to obtain a classification importance score for
820 each gene, indicative of the extent to which this gene's expression can discriminate HRD and HR-
821 proficient samples. The pipeline consisted of following steps:

822 (1) A weighted correlation network analysis (WGCNA)(72) method was employed to extract a
823 gene co-expression graph involving all 228 genes in the HRD transcriptional signature across the
824 TCGA-BRCA cohort;

825 (2) A GAT-based feature extractor was applied to the high dimensional embeddings of the gene
826 co-expression graphs which integrates information from neighbouring domains;

827 (3) A simple prediction module for the classification task was implemented;

828 (4) A gradient-based method was used to calculate gene importance scores as post-hoc explanations
829 for model behaviour(120).

830 This pipeline was implemented based on the Pytorch(121) and Pytorch Geometric library(122). An
831 Adam optimizer was used with batch size 8 and an initial learning rate of 0.002. Linear learning
832 rate decay and early stopping were applied to avoid overfitting. A threshold of 0.4 was applied to
833 identify 'important' genes for HRD and HR-proficiency classification.

834 **Cell-cell interaction analysis**

835 Differential patterns of cell-cell interactivity within the tumour microenvironment of HRD and HR-
836 proficient cells were analysed using CellphoneDB(86), which was applied to the Qian et al. and
837 Bassez et al. cohorts(71). Tumour cells were labelled as HRD if they displayed a positive HRD
838 score. CellphoneDB was conducted within a Conda virtual Python environment, with default
839 parameters applied.

840 **Statistical analysis**

841 Groups were compared using a two-sided Student's t test, Wilcoxon rank-sum test or ANOVA, as
842 appropriate. P-values were adjusted for multiple testing where appropriate using the Benjamini-
843 Hochberg method. Graphs were generated using the ggplot2 and ggpubr R packages.

844

845 **DECLARATIONS**

846 **Ethics, consent and permissions**

847 All data employed in this study comply with ethical regulations, with approval and informed
848 consent for collection and sharing already obtained by the relevant consortia where the data were
849 obtained from (TCGA, ICGC).

850 **Consent for publication**

851 Not applicable.

852 **Availability of data and materials**

853 The results published here are in part based upon data generated by the TCGA Research Network
854 (https://www.cancer.gov/tcga), ICGC (https://dcc.icgc.org/), or deposited at cBioPortal
855 (https://www.cbioportal.org/), CCLE (https://sites.broadinstitute.org/ccle/) and GEO
856 (https://www.ncbi.nlm.nih.gov/geo/).

857 All code developed for the purpose of this study can be found at the following repository:
858 https://github.com/secrierlab/MultiscaleHRD

859 **Competing interests**

860 The authors declare that they have no competing interests.

861 **Funding**

862 DHJ was supported by an MRC DTP grant (MR/N013867/1). MS and SP were supported by a
863 UKRI Future Leaders Fellowship (MR/T042184/1). Work in MS's lab was supported by a BBSRC
864 equipment grant (BB/R01356X/1) and a Wellcome Institutional Strategic Support Fund
865 (204841/Z/16/Z). JF acknowledges the National Institute for Health Research University College
866 London Hospitals Biomedical Research Centre.

867 **Authors' contributions**

868 MS designed and supervised the study. DHJ developed the mutational and transcriptional
869 classifiers and performed all analyses in bulk and single cell datasets. SP developed and applied
870 the graph neural network classifier. JF helped co-supervise DHJ and provided valuable feedback
871 on the analyses performed in the study.

872

873 **REFERENCES**

874 1. Ciccia A, Elledge SJ. The DNA Damage Response: Making It Safe to Play with Knives.
875 Vol. 40, Molecular Cell. 2010. p. 179–204.
876 2. Lord CJ, Ashworth A. The DNA damage response and cancer therapy. Vol. 481, Nature.
877 2012. p. 287–94.
878 3. Ceccaldi R, Liu JC, Amunugama R, Hajdu I, Primack B, Petalcorin MIR, et al.
879 Homologous-recombination-deficient tumours are dependent on Polθ -mediated repair.
880 Nature. 2015 Feb 12;518(7538):258–62.
881 4. O'Connor MJ. Targeting the DNA Damage Response in Cancer. Vol. 60, Molecular Cell.
882 Cell Press; 2015. p. 547–60.
883 5. Fong PC, Boss DS, Yap TA, Tutt A, Wu P, Mergui-Roelvink M, et al. Inhibition of
884 Poly(ADP-Ribose) Polymerase in Tumors from BRCA Mutation Carriers . New England
885 Journal of Medicine. 2009 Jul 9;361(2):123–34.
886 6. Zatreanu D, Robinson HMR, Alkhatib O, Boursier M, Finch H, Geo L, et al. Polθ
887 inhibitors elicit BRCA-gene synthetic lethality and target PARP inhibitor resistance. Nat
888 Commun. 2021 Dec 1;12(1).
889 7. Zhou J, Gelot C, Pantelidou C, Li A, Yücel H, Davis RE, et al. A first-in-class polymerase
890 theta inhibitor selectively targets homologous-recombination-deficient tumors. Nat
891 Cancer. 2021 Jun 1;2(6):598–610.
892 8. Moore K, Colombo N, Scambia G, Kim BG, Oaknin A, Friedlander M, et al. Maintenance
893 Olaparib in Patients with Newly Diagnosed Advanced Ovarian Cancer. New England
894 Journal of Medicine. 2018 Dec 27;379(26):2495–505.
895 9. Tutt ANJ, Garber JE, Kaufman B, Viale G, Fumagalli D, Rastogi P, et al. Adjuvant
896 Olaparib for Patients with BRCA1 - or BRCA2 -Mutated Breast Cancer . New England
897 Journal of Medicine. 2021 Jun 24;384(25):2394–405.
898 10. Litton JK, Rugo HS, Ettl J, Hurvitz SA, Gonçalves A, Lee KH, et al. Talazoparib in
899 Patients with Advanced Breast Cancer and a Germline BRCA Mutation . New England
900 Journal of Medicine. 2018 Aug 23;379(8):753–63.
901 11. Davies H, Glodzik D, Morganella S, Yates LR, Staaf J, Zou X, et al. HRDetect is a
902 predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat Med.
903 2017 Apr 1;23(4):517–25.
904 12. Nguyen L, W. M. Martens J, Van Hoeck A, Cuppen E. Pan-cancer landscape of
905 homologous recombination deficiency. Nat Commun. 2020 Dec 1;11(1).
906 13. Ladan MM, van Gent DC, Jager A. Homologous recombination deficiency testing for
907 brca-like tumors: The road to clinical validation. Vol. 13, Cancers. MDPI AG; 2021. p. 1–
908 23.

14. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin A V., et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415–21.
15. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, et al. The life history of 21 breast cancers. Cell. 2012 May 25;149(5):994–1007.
16. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature. 2016 May 2;534(7605):47–54.
17. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020 Feb 6;578(7793):94–101.
18. Macintyre G, Goranova TE, De Silva D, Ennis D, Piskorz AM, Eldridge M, et al. Copy number signatures and mutational processes in ovarian carcinoma. Nat Genet. 2018 Sep 1;50(9):1262–70.
19. Wang S, Li H, Song M, Tao Z, Wu T, He Z, et al. Copy number signature analysis tool and its application in prostate cancer reveals distinct mutational processes and clinical outcomes. PLoS Genet. 2021 May 4;17(5 May).
20. Drews RM, Hernando B, Tarabichi M, Haase K, Lesluyes T, Smith PS, et al. A pan-cancer compendium of chromosomal instability. Nature. 2022 Jun 30;606(7916):976–83.
21. Steele CD, Abbasi A, Islam SMA, Bowes AL, Khandekar A, Haase K, et al. Signatures of copy number alterations in human cancer. Nature. 2022 Jun 30;606(7916):984–91.
22. Marquard AM, Eklund AC, Joshi T, Krzystanek M, Favero F, Wang ZC, et al. Pan-cancer analysis of genomic scar signatures associated with homologous recombination deficiency suggests novel indications for existing cancer drugs. Biomark Res. 2015 Dec;3(1).
23. Melinda LT, Kirsten MT, Julia R, Bryan H, Gordon BM, Kristin CJ, et al. Homologous recombination deficiency (hrd) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple-negative breast cancer. Clinical Cancer Research. 2016 Aug 1;22(15):3764–73.
24. Birkbak NJ, Wang ZC, Kim JY, Eklund AC, Li Q, Tian R, et al. Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. Cancer Discov. 2012 Apr;2(4):366–75.
25. Popova T, Manié E, Rieunier G, Caux-Moncoutier V, Tirapo C, Dubois T, et al. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. Cancer Res. 2012 Nov 1;72(21):5454–62.
26. Abkevich V, Timms KM, Hennessy BT, Potter J, Carey MS, Meyer LA, et al. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. Br J Cancer. 2012 Nov 6;107(10):1776–82.
27. Gulhan DC, Lee JJK, Melloni GEM, Cortés-Ciriano I, Park PJ. Detecting the mutational signature of homologous recombination deficiency in clinical samples. Nat Genet. 2019 May 1;51(5):912–9.
28. Batalini F, Gulhan DC, Mao V, Tran A, Polak M, Xiong N, et al. Mutational Signature 3 Detected from Clinical Panel Sequencing is Associated with Responses to Olaparib in Breast and Ovarian Cancers. Clinical Cancer Research. 2022 Nov 1;28(21):4714–23.
29. Dias MP, Moser SC, Ganesan S, Jonkers J. Understanding and overcoming resistance to PARP inhibitors in cancer therapy. Vol. 18, Nature Reviews Clinical Oncology. Nature Research; 2021. p. 773–91.
30. Noordermeer SM, Adam S, Setiaputra D, Barazas M, Pettitt SJ, Ling AK, et al. The shieldin complex mediates 53BP1-dependent DNA repair. Nature. 2018;560(7716):117–21.

957 31. Severson TM, Wolf DM, Yau C, Peeters J, Wehkam D, Schouten PC, et al. The
958 BRCA1ness signature is associated significantly with response to PARP inhibitor
959 treatment versus control in the I-SPY 2 randomized neoadjuvant setting. Breast Cancer
960 Research. 2017 Aug 25;19(1).
961 32. Peng G, Lin CCJ, Mo W, Dai H, Park YY, Kim SM, et al. Genome-wide transcriptome
962 profiling of homologous recombination DNA repair. Nat Commun. 2014 Feb 20;5.
963 33. Daemen A, Wolf DM, Korkola JE, Griffith OL, Frankum JR, Brough R, et al. Cross-
964 platform pathway-based analysis identifies markers of response to the PARP inhibitor
965 olaparib. Breast Cancer Res Treat. 2012 Sep;135(2):505–17.
966 34. Carter SL, Eklund AC, Kohane IS, Harris LN, Szallasi Z. A signature of chromosomal
967 instability inferred from gene expression profiles predicts clinical outcome in multiple
968 human cancers. Nat Genet. 2006 Sep 4;38(9):1043–8.
969 35. Pan JW, Ng PS, Mamduh M, Zabidi A, Fatin PN, Teo JY, et al. Gene signature for
970 predicting homologous recombination deficiency in triple-negative breast cancer. BioRxiv
971 [Internet]. 2022; Available from: https://doi.org/10.1101/2022.06.08.495296
972 36. Sunada S, Nakanishi A, Miki Y. Crosstalk of DNA double-strand break repair pathways in
973 poly(ADP-ribose) polymerase inhibitor treatment of breast cancer susceptibility gene 1/2-
974 mutated cancer. Vol. 109, Cancer Science. Blackwell Publishing Ltd; 2018. p. 893–9.
975 37. Prakash R, Zhang Y, Feng W, Jasin M. Homologous recombination and human health:
976 The roles of BRCA1, BRCA2, and associated proteins. Cold Spring Harb Perspect Biol.
977 2015;7(4).
978 38. Pusztai L, Yau C, Wolf DM, Han HS, Du L, Wallace AM, et al. Durvalumab with olaparib
979 and paclitaxel for high-risk HER2-negative stage II/III breast cancer: Results from the
980 adaptively randomized I-SPY2 trial. Cancer Cell. 2021 Jul 12;39(7):989-998.e5.
981 39. Hwang T, Reh S, Dunbayev Y, Zhong Y, Takata Y, Shen J, et al. Defining the mutation
982 signatures of DNA polymerase θ in cancer genomes. NAR Cancer. 2020 Sep 1;2(3).
983 40. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. deconstructSigs:
984 Delineating mutational processes in single tumors distinguishes DNA repair deficiencies
985 and patterns of carcinoma evolution. Genome Biol. 2016 Feb 22;17(1).
986 41. Perez-Villatoro F, Oikkonen J, Casado J, Chernenko A, Gulhan DC, Tumiati M, et al.
987 Optimized detection of homologous recombination deficiency improves the prediction of
988 clinical outcomes in cancer. NPJ Precis Oncol. 2022 Dec 1;6(1).
989 42. Kan Z, Ding Y, Kim J, Jung HH, Chung W, Lal S, et al. Multi-omics profiling of younger
990 Asian breast cancers reveals distinctive molecular signatures. Nat Commun. 2018 Dec
991 1;9(1).
992 43. Telli ML, Stover DG, Loi S, Aparicio S, Carey LA, Domchek SM, et al. Homologous
993 recombination deficiency and host anti-tumor immunity in triple-negative breast cancer.
994 Vol. 171, Breast Cancer Research and Treatment. Springer New York LLC; 2018. p. 21–
995 31.
996 44. Takaya H, Nakai H, Takamatsu S, Mandai M, Matsumura N. Homologous recombination
997 deficiency status-based classification of high-grade serous ovarian carcinoma. Sci Rep.
998 2020 Dec 1;10(1).
999 45. Schrempf A, Slyskova J, Loizou JI. Targeting the DNA Repair Enzyme Polymerase θ in
1000 Cancer Therapy. Vol. 7, Trends in Cancer. Cell Press; 2021. p. 98–111.
1001 46. Patterson-Fortin J, D'Andrea AD. Exploiting the microhomology-mediated end-joining
1002 pathway in cancer therapy. Vol. 80, Cancer Research. American Association for Cancer
1003 Research Inc.; 2020. p. 4593–600.

1004 47. Wiecek AJ, Cutty SJ, Kornai D, Parreno-Centeno M, Gourmet LE, Tagliazucchi GM, et al.
1005 Genomic hallmarks and therapeutic implications of G0 cell cycle arrest in cancer. Genome
1006 Biol. 2023 Dec 1;24(1).
1007 48. Dominguez-Sola D, Gautier J. MYC and the control of DNA replication. Cold Spring
1008 Harb Perspect Med. 2014;4(6).
1009 49. McAlpine JN, Porter H, Köbel M, Nelson BH, Prentice LM, Kalloger SE, et al. BRCA1
1010 and BRCA2 mutations correlate with TP53 abnormalities and presence of immune cell
1011 infiltrates in ovarian high-grade serous carcinoma. Modern Pathology. 2012
1012 May;25(5):740–50.
1013 50. Greenblatt MS, Chappuis PO, Bond JP, Hamel N, Foulkes WD. TP53 Mutations in Breast
1014 Cancer Associated with BRCA1 or BRCA2 Germ-line Mutations: Distinctive Spectrum
1015 and Structural Distribution 1 [Internet]. Vol. 61, CANCER RESEARCH. 2001. Available
1016 from: http://metalab.unc.edu/dnam/mainpage.html.
1017 51. Takaku M, Grimm SA, Wade PA. GATA3 in breast cancer: Tumor suppressor or
1018 oncogene? Vol. 16, Gene Expression. Cognizant Communication Corporation; 2015. p.
1019 163–8.
1020 52. Cohen H, Ben-Hamo R, Gidoni M, Yitzhaki I, Kozol R, Zilberberg A, et al. Shift in
1021 GATA3 functions, and GATA3 mutations, control progression and clinical presentation in
1022 breast cancer. Breast Cancer Research. 2014;16(1).
1023 53. Ansari-Pour N, Zheng Y, Yoshimatsu TF, Sanni A, Ajani M, Reynier JB, et al. Whole-
1024 genome analysis of Nigerian patients with breast cancer reveals ethnic-driven somatic
1025 evolution and distinct genomic subtypes. Nat Commun. 2021 Dec 1;12(1).
1026 54. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal
1027 Patterns of Selection in Cancer and Somatic Tissues. Cell. 2017 Nov 16;171(5):1029-
1028 1041.e21.
1029 55. Heijink AM, Talens F, Jae LT, van Gijn SE, Fehrmann RSN, Brummelkamp TR, et al.
1030 BRCA2 deficiency instigates cGAS-mediated inflammatory signaling and confers
1031 sensitivity to tumor necrosis factor-alpha-mediated cytotoxicity. Nat Commun. 2019 Dec
1032 1;10(1).
1033 56. Pawlyn C, Loehr A, Ashby C, Tytarenko R, Deshpande S, Sun J, et al. Loss of
1034 heterozygosity as a marker of homologous repair deficiency in multiple myeloma: A role
1035 for PARP inhibition? Leukemia. 2018 Jul 1;32(7):1561–6.
1036 57. Gruber JJ, Afghahi A, Timms K, DeWees A, Gross W, Aushev VN, et al. A phase II study
1037 of talazoparib monotherapy in patients with wild-type BRCA1 and BRCA2 with a
1038 mutation in other homologous recombination genes. Nat Cancer. 2022 Oct 1;3(10):1181–
1039 91.
1040 58. Dillon KM, Bekele RT, Sztupinszki Z, Hanlon T, Rafiei S, Szallasi Z, et al. PALB2 or
1041 BARD1 loss confers homologous recombination deficiency and PARP inhibitor sensitivity
1042 in prostate cancer. NPJ Precis Oncol. 2022 Dec 1;6(1).
1043 59. Zhang M, Liu G, Xue F, Edwards R, Sood AK, Zhang W, et al. Copy number deletion of
1044 RAD50 as predictive marker of BRCAness and PARP inhibitor response in BRCA wild
1045 type ovarian cancer. Gynecol Oncol. 2016 Apr 1;141(1):57–64.
1046 60. Chang HHY, Pannunzio NR, Adachi N, Lieber MR. Non-homologous DNA end joining
1047 and alternative pathways to double-strand break repair. Vol. 18, Nature Reviews
1048 Molecular Cell Biology. Nature Publishing Group; 2017. p. 495–506.
1049 61. Miyagawa K, Tsuruga T, Kinomura A, Usui K, Katsura M, Tashiro S, et al. A role for
1050 RAD54B in homologous recombination in human cells. EMBO J. 2002;21:175–80.

1051 62. Prakash R, Sandoval T, Morati F, Zagelbaum JA, Lim PX, White T, et al. Distinct
1052 pathways of homologous recombination controlled by the SWS1–SWSAP1–SPIDR
1053 complex. Nat Commun. 2021 Dec 1;12(1).
1054 63. Pearl LH, Schierz AC, Ward SE, Al-Lazikani B, Pearl FMG. Therapeutic opportunities
1055 within the DNA damage response. Nat Rev Cancer. 2015 Feb 24;15(3):166–80.
1056 64. Mandal J, Mandal P, Wang TL, Shih IM. Treating ARID1A mutated cancers by
1057 harnessing synthetic lethality and DNA damage response. Vol. 29, Journal of Biomedical
1058 Science. BioMed Central Ltd; 2022.
1059 65. Badia-I-Mompel P, Vélez Santiago J, Braunger J, Geiss C, Dimitrov D, Müller-Dott S, et
1060 al. decoupleR: ensemble of computational methods to infer biological activities from
1061 omics data. Bioinformatics Advances. 2022;2(1).
1062 66. Mehibel M, Xu Y, Li CG, Moon EJ, Thakkar KN, Diep AN, et al. Eliminating hypoxic
1063 tumor cells improves response to PARP inhibitors in homologous recombination–deficient
1064 cancer models. Journal of Clinical Investigation. 2021 Jun 1;131(11).
1065 67. Chan N, Pires IM, Bencokova Z, Coackley C, Luoto KR, Bhogal N, et al. Contextual
1066 synthetic lethality of cancer cell kill based on the tumor microenvironment. Cancer Res.
1067 2010 Oct 15;70(20):8045–54.
1068 68. Wolff M, Kosyna FK, Dunst J, Jelkmann W, Depping R. Impact of hypoxia inducible
1069 factors on estrogen receptor expression in breast cancer cells. Arch Biochem Biophys.
1070 2017 Jan 1;613:23–30.
1071 69. Chu T, Wang Z, Pe'er D, Danko CG. Cell type and gene expression deconvolution with
1072 BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA
1073 sequencing in oncology. Nat Cancer. 2022 Apr 1;3(4):505–17.
1074 70. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat
1075 Commun. 2015 Dec 4;6.
1076 71. Qian J, Olbrecht S, Boeckx B, Vos H, Laoui D, Etlioglu E, et al. A pan-cancer blueprint of
1077 the heterogeneous tumor microenvironment revealed by single-cell profiling. Cell Res.
1078 2020 Sep 1;30(9):745–62.
1079 72. Rezaie N, Reese F, Mortazavi A. PyWGCNA: A Python package for weighted gene co-
1080 expression network analysis. BioRxiv [Internet]. 2022; Available from:
1081 https://doi.org/10.1101/2022.08.22.504852
1082 73. Wang Q, Sun Z, Xia W, Sun L, Du Y, Zhang Y, et al. Role of USP13 in physiology and
1083 diseases. Vol. 9, Frontiers in Molecular Biosciences. Frontiers Media S.A.; 2022.
1084 74. Kim W, Zhao F, Gao H, Qin S, Hou J, Deng M, et al. USP13 regulates the replication
1085 stress response by deubiquitinating TopBP1. DNA Repair (Amst). 2021 Apr 1;100.
1086 75. Singh KK, Ayyasamy V, Owens KM, Koul MS, Vujcic M. Mutations in mitochondrial
1087 DNA polymerase-γ promote breast tumorigenesis. J Hum Genet. 2009 Sep;54(9):516–24.
1088 76. Copeland WC. Defects of mitochondrial DNA replication. In: Journal of Child Neurology.
1089 SAGE Publications Inc.; 2014. p. 1216–24.
1090 77. Koldobskiy MA, Chakraborty A, Werner JK, Snowman AM, Juluri KR, Scott Vandiver
1091 M, et al. p53-mediated apoptosis requires inositol hexakisphosphate kinase-2. PNAS
1092 [Internet]. 2010;107(49):20947–51. Available from:
1093 www.pnas.org/cgi/doi/10.1073/pnas.1015671107
1094 78. Rao F, Cha J, Xu J, Xu R, Vandiver MS, Tyagi R, et al. Inositol Pyrophosphates Mediate
1095 the DNA-PK/ATM-p53 Cell Death Pathway by Regulating CK2 Phosphorylation of
1096 Tti1/Tel2. Mol Cell. 2014 Apr 10;54(1):119–32.
1097 79. Vilas CK, Emery LE, Denchi EL, Miller KM. Caught with One's Zinc Fingers in the
1098 Genome Integrity Cookie Jar. Vol. 34, Trends in Genetics. Elsevier Ltd; 2018. p. 313–25.

1099 80. Singh JK, van Attikum H. DNA double-strand break repair: Putting zinc fingers on the
1100 sore spot. Vol. 113, Seminars in Cell and Developmental Biology. Elsevier Ltd; 2021. p.
1101 65–74.
1102 81. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The
1103 Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.
1104 Nature. 2012 Mar 29;483(7391):603–7.
1105 82. Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, et al. Single-cell RNA-seq
1106 enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat
1107 Commun. 2017 May 5;8.
1108 83. Feeley LP, Mulligan AM, Pinnaduwage D, Bull SB, Andrulis IL. Distinguishing luminal
1109 breast cancer subtypes by Ki67, progesterone receptor or TP53 status provides prognostic
1110 information. Modern Pathology. 2014;27(4):554–61.
1111 84. Cheang MCU, Chia SK, Voduc D, Gao D, Leung S, Snider J, et al. Ki67 index, HER2
1112 status, and prognosis of patients with luminal B breast cancer. J Natl Cancer Inst. 2009
1113 May;101(10):736–50.
1114 85. Bassez A, Vos H, Van Dyck L, Floris G, Arijs I, Desmedt C, et al. A single-cell map of
1115 intratumoral changes during anti-PD1 treatment of patients with breast cancer. Nat Med.
1116 2021 May 1;27(5):820–32.
1117 86. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. CellPhoneDB: Inferring
1118 cell-cell communication from combined expression of multi-subunit ligand-receptor
1119 complexes [Internet]. 2020. Available from: http://www.cellphonedb.org/
1120 87. Irie A, Kontani K, Kihara M, Liu D, Shirato Y, Seki M, et al. Galectin-9 as a Prognostic
1121 Factor with Antimetastatic Potential in Breast Cancer. Clin Cancer Res [Internet].
1122 2005;11(8):2962–8. Available from: www.aacrjournals.org
1123 88. Bakhoum SF, Ngo B, Laughney AM, Cavallo JA, Murphy CJ, Ly P, et al. Chromosomal
1124 instability drives metastasis through a cytosolic DNA response. Nature. 2018 Jan
1125 25;553(7689):467–72.
1126 89. Yoon HK, Kim TH, Park SG, Jung H, Quan X, Park SJ, et al. Effect of anthracycline and
1127 taxane on the expression of programmed cell death ligand-1 and galectin-9 in triple-
1128 negative breast cancer. Pathol Res Pract. 2018 Oct 1;214(10):1626–31.
1129 90. Lv Y, Ma X, Ma Y, Du Y, Feng J. A new emerging target in cancer immunotherapy:
1130 Galectin-9 (LGALS9). Genes and Diseases. Chongqing University; 2022.
1131 91. Ben-David U, Siranosian B, Ha G, Tang H, Oren Y, Hinohara K, et al. Genetic and
1132 transcriptional evolution alters cancer cell line drug response. Nature. 2018 Aug
1133 16;560(7718):325–30.
1134 92. Bhin J, Paes Dias M, Gogola E, Rolfs F, Piersma SR, de Bruijn R, et al. Multi-omics
1135 analysis reveals distinct non-reversion mechanisms of PARPi resistance in BRCA1- versus
1136 BRCA2-deficient mammary tumors. Cell Rep. 2023 May 30;42(5).
1137 93. França GS, Baron M, Pour M, King BR, Rao A, Misirlioglu S, et al. Drug-induced
1138 adaptation along a resistance continuum in cancer cells. BioRxiv [Internet]. 2022;
1139 Available from: https://doi.org/10.1101/2022.06.21.496830
1140 94. Funnell T, O'Flanagan CH, Williams MJ, McPherson A, McKinney S, Kabeer F, et al.
1141 Single-cell genomic variation induced by mutational processes in cancer. Nature. 2022;
1142 95. Budczies J, Kluck K, Beck S, Ourailidis I, Allgäuer M, Menzel M, et al. Homologous
1143 recombination deficiency is inversely correlated with microsatellite instability and
1144 identifies immunologically cold tumors in most cancer types. Journal of Pathology:
1145 Clinical Research. 2022 Jul 1;8(4):371–82.
1146 96. Pellegrino B, Musolino A, Llop-Guevara A, Serra V, De Silva P, Hlavata Z, et al.
1147 Homologous Recombination Repair Deficiency and the Immune Response in Breast

Cancer: A Literature Review. Vol. 13, Translational Oncology. Neoplasia Press, Inc.; 2020. p. 410–22.
97. Sumitani N, Ishida K, Sawada K, Kimura T, Kaneda Y, Nimura K. Identification of Malignant Cell Populations Associated with Poor Prognosis in High-Grade Serous Ovarian Cancer Using Single-Cell RNA Sequencing. Cancers (Basel). 2022 Aug 1;14(15).
98. Launonen IM, Lyytikäinen N, Casado J, Anttila EA, Szabó A, Haltia UM, et al. Single-cell tumor-immune microenvironment of BRCA1/2 mutated high-grade serous ovarian cancer. Nat Commun. 2022 Dec 1;13(1).
99. Tietscher S, Wagner J, Anzeneder T, Langwieder C, Rees M, Sobottka B, et al. A comprehensive single-cell map of T cell exhaustion-associated immune environments in human breast cancer. Nat Commun [Internet]. 2023 Jan 6;14(1):98. Available from: https://www.nature.com/articles/s41467-022-35238-w
100. Vikas P, Borcherding N, Chennamadhavuni A, Garje R. Therapeutic Potential of Combining PARP Inhibitor and Immunotherapy in Solid Tumors. Vol. 10, Frontiers in Oncology. Frontiers Media S.A.; 2020.
101. Peyraud F, Italiano A. Combined parp inhibition and immune checkpoint therapy in solid tumors. Vol. 12, Cancers. MDPI AG; 2020. p. 1–28.
102. Hong C, Schubert M, Tijhuis AE, Requesens M, Roorda M, van den Brink A, et al. cGAS–STING drives the IL-6-dependent survival of chromosomally instable cancers. Nature. 2022 Jul 14;607(7918):366–73.
103. Zhang J, Bajari R, Andric D, Gerthoffert F, Lepsa A, Nahal-Bose H, et al. The International Cancer Genome Consortium Data Portal. Nat Biotechnol. 2019 Apr 15;37(4):367–9.
104. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016 May 5;44(8):e71.
105. Valieris R, Amaro L, de Toledo Osório CAB, Bueno AP, Mitrowsky RAR, Carraro DM, et al. Deep learning predicts underlying features on pathology images with therapeutic relevance for breast and gastric cancer. Cancers (Basel). 2020 Dec 1;12(12):1–12.
106. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012 May;2(5):401–4.
107. Wiecek AJ, Cutty SJ, Kornai D, Parreno-Centeno M, Gourmet LE, Tagliazucchi GM, et al. Genomic hallmarks and therapeutic implications of cancer cell quiescence Running title: Hallmarks of cancer quiescence and therapeutic implications. BioRxiv [Internet]. 2022; Available from: https://doi.org/10.1101/2021.11.12.468410
108. Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun. 2017 May 5;8.
109. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun 24;184(13):3573-3587.e29.
110. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models [Internet]. Available from: http://cran-logs.rstudio.com
111. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: The Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019 Jan 8;47(D1):D941–7.

112. Schubert M, Klinger B, Klünemann M, Sieber A, Uhlitz F, Sauer S, et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat Commun. 2018 Dec 1;9(1).
113. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014 Dec 5;15(12).
114. Buffa FM, Harris AL, West CM, Miller CJ. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. Br J Cancer. 2010;102(2):428–35.
115. Bhandari V, Hoey C, Liu LY, Lalonde E, Ray J, Livingstone J, et al. Molecular landmarks of tumor hypoxia across cancer types. Nat Genet. 2019 Feb 1;51(2):308–18.
116. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Softw [Internet]. 2008;28(5). Available from: http://www.jstatsoft.org/
117. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent [Internet]. Vol. 33, JSS Journal of Statistical Software. 2010. Available from: http://www.jstatsoft.org/
118. Ulgen E, Ozisik O, Sezerman OU. PathfindR: An R package for comprehensive identification of enriched pathways in omics data through active subnetworks. Front Genet. 2019 Sep 1;10(SEP).
119. Kuleshov M V., Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016 Jul 8;44(1):W90–7.
120. Xing X, Yang F, Li H, Zhang J, Zhao Y, Gao M, et al. Multi-level attention graph neural network based on co-expression gene modules for disease diagnosis and prognosis. Bioinformatics. 2022 Apr 15;38(8):2178–86.
121. Paszke A, Gross S, Massa F, Lerer A, Bradbury Google J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in neural information proceessing systems. 2019;32.
122. Fey M, Lenssen JE. Fast Graph Representation Learning with PyTorch Geometric. ArXiv [Internet]. 2019 Mar 6; Available from: http://arxiv.org/abs/1903.02428

FIGURES

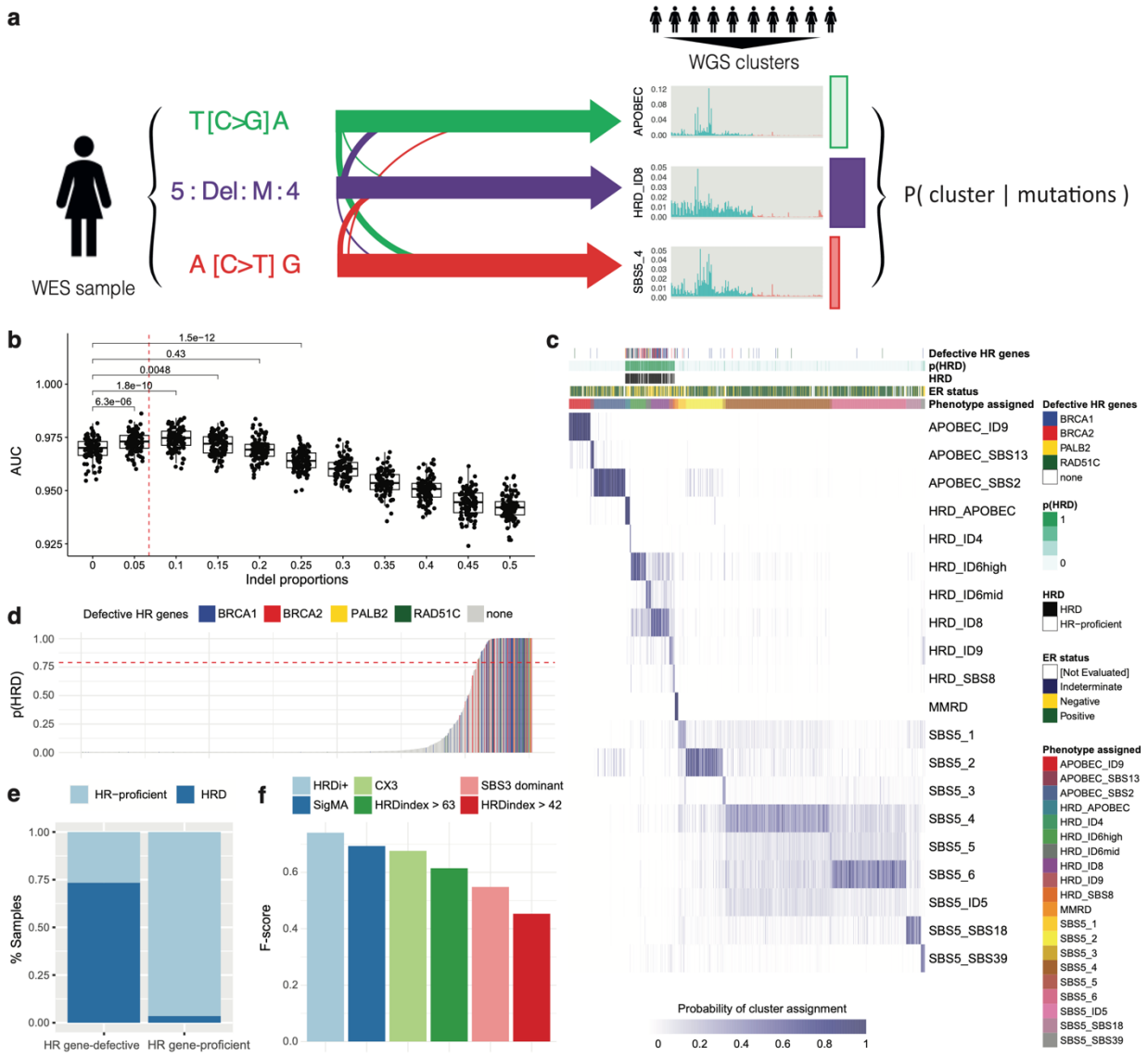

Figure 1. Evaluating HRD in exome sequenced breast cancers. (a) Workflow for HRD classification of an exome-sequenced breast cancer sample. Each sample contains a profile of mutations, each of which has a probabilistic association with each of the 20 signature phenotypes, defined by a representative signature profile inferred from WGS data. The mutational profile is collated to calculate the probability of assignment of the respective sample to each of the 20 clusters. (b) Simulation analysis of SBS3-enrichment classification of ICGC samples downsampled to 50 mutational events constrained to varying indel proportions. Adding a small percentage of indels is sufficient to improve classification. AUC = Area Under the ROC Curve for SBS3 enrichment classification. The dotted red line represents the mean proportion of indel events in the TCGA-BRCA cohort. (c) Classification of 968 exome sequenced breast cancer samples from TCGA. The heat map indicates the probability of each sample (column) being assigned to each signature phenotype (rows). Samples are annotated by ER status and HR gene defects. The value p(HRD) is the sum of probabilities of assignment across the seven HRD-associated phenotypes. The label 'Phenotype assigned' refers to the phenotype to which the respective sample has the highest probability of assignment. (d) Summary of HRD cluster assignment probabilities across the TCGA-BRCA cohort. Samples with a total probability of HRD assignment greater than 0.79

1243 (as shown by the dotted red line) would be assigned as HRD, whereas the rest would be deemed HR-
1244 proficient. (e) HRD classification of HR gene-defective and -positive samples in the TCGA-BRCA
1245 cohort. 'HRD' refers to samples with a probability of HRD assignment greater than 0.79. (f) F-score
1246 comparisons of HRD classifiers for exomes. 'HRDi+' refers to the classifier developed in this study.
1247 The HRD index is presented using cut-offs of 42 and 63.

1248

1249

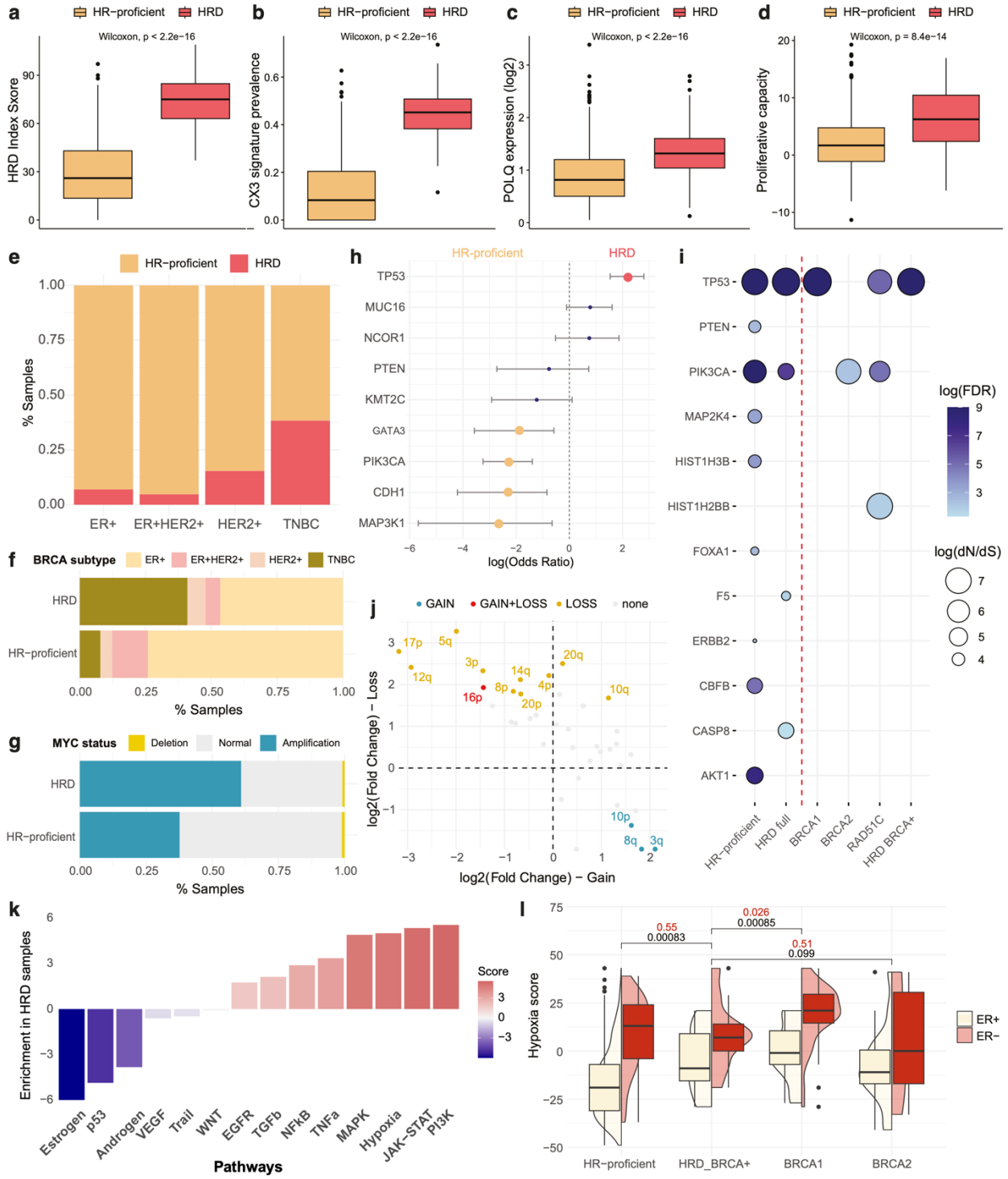

Figure 2. Genomic and transcriptional hallmarks of HRD. Association between HRD status and (a) the Myriad HRD index score, (b) contribution of the CX3 copy number signature, (c) POLQ expression, and (d) a transcriptional measurement of proliferation/cell cycle arrest capacity. (e) Enrichment of HRD across breast cancer subtypes. (f) Enrichment of breast cancer subtypes across HRD status. (g) Association between HRD status and amplification of MYC. (h) Enrichment/depletion of somatic nonsynonymous mutations in key cancer genes in HRD and HR-proficient breast cancer samples. A positive log(Odds Ratio) indicates enrichment in HRD samples. (i) Positive selection of cancer genes in HR-proficient, all HRD, BRCA1-defective, BRCA2-defective, RAD51C-defective, and HRD

1259 BRCA+ breast cancer samples. Circle size indicates the strength of positive selection according to the
1260 dN/dS ratio. (j) Comparison of chromosome arm loss and gain events between HRD and HR-proficient
1261 breast cancer samples. Positive values indicate enrichment in HRD against HR-proficient samples,
1262 whilst the x and y axes indicate enrichment for chromosome arm gains and losses respectively. (k)
1263 Results of differential pathway activity analysis between HRD and HR-proficient breast cancer samples
1264 across 14 signalling pathways ordered by the Normalised Enrichment Score (NES). Positive scores
1265 indicate pathway enrichment in HRD samples. (l) Comparison of hypoxia scores in the TCGA-BRCA
1266 cohort according to the Buffa transcriptional signature across HRD/BRCA-defect categories, split by
1267 ER status. P-values refer to Wilcoxon testing between each group and the HRD-BRCA+ group, tested
1268 across all samples (black) and ER-negative samples only (red).

1269

1270

1271

1272

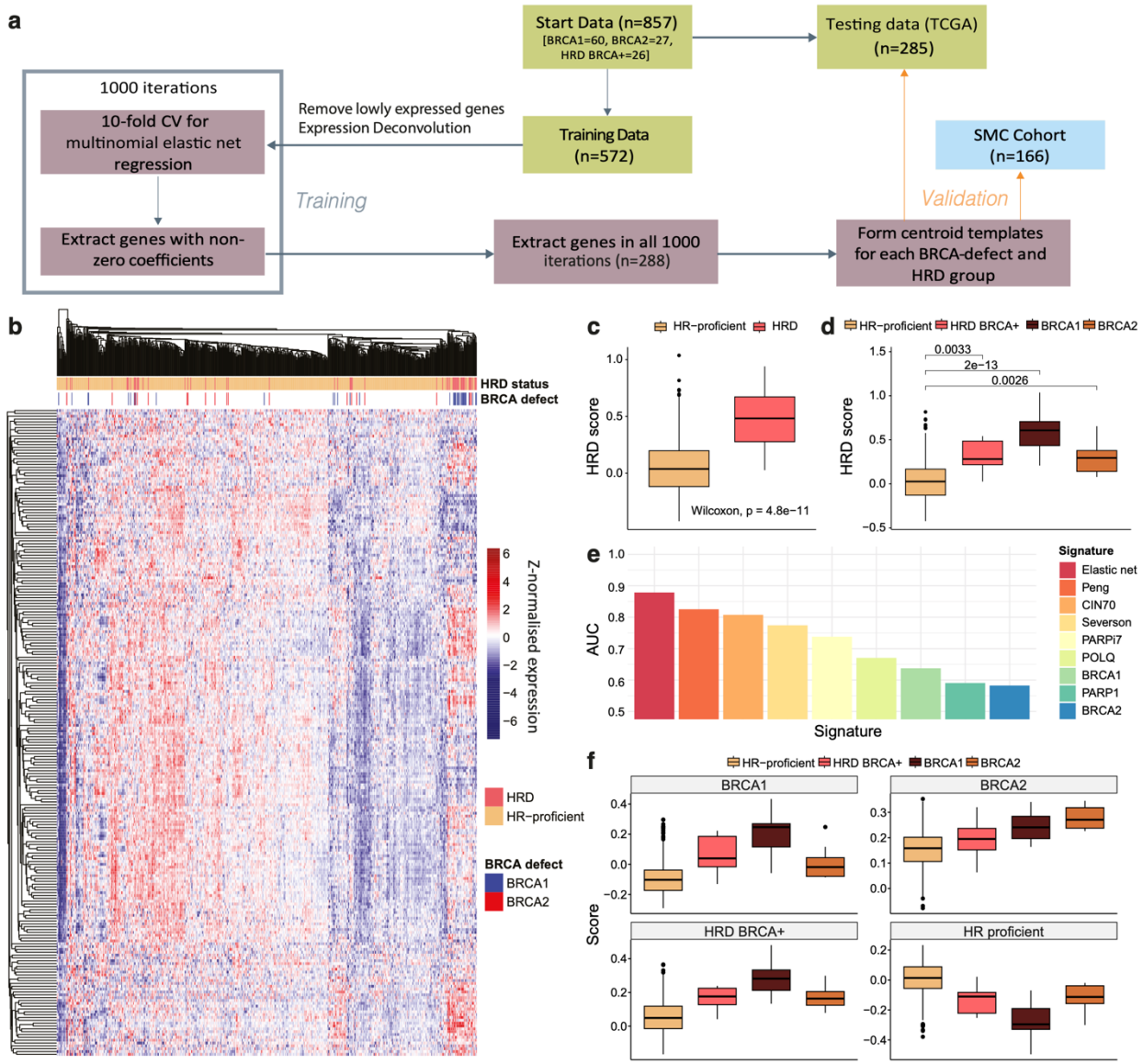

Figure 3. Development and validation of a BRCA defect type-specific HRD transcriptional signature. (a) Workflow for transcriptional signature development. Data is split into training and testing cohorts. The training data undergoes expression deconvolution to extract a cancer cell-specific signal, using the Qian et al. single-cell RNA-seq cohort as a reference, and genes that are lowly expressed in this dataset are removed. Processed training data undergoes 1000 iterations of 10-fold cross validation of elastic net regression, and a signature is formed from the 228 genes selected in every iteration. Centroid templates are formed for HRD/HR-proficient and BRCA-type HRD groups from the 228 genes across the training cohort, and scores for testing and validation cohorts are calculated by correlating the new sample against each template. (b) Summary of the 228-gene HRD transcriptional signature profiles across the TCGA training set. The HRD status assignment is annotated along with BRCA1/2 defects. (c-d) Comparison of HRD scores calculated using the transcriptional signature between c) HRD vs HR-proficient and d) HRD/BRCA-defect groups. (e) Comparison of HRD transcriptional signatures and gene expression markers for predicting HRD status in the TCGA training set, measured by AUC. 'Elastic net' refers to the 228-gene transcriptional signature presented in this study. 'Peng', 'CIN70', 'Severson', and 'PARPi7' refer to alternative transcriptional signatures as described in the Methods. POLQ, BRCA1, PARP1, and BRCA2 are gene expression markers. (f) Comparison of HRD/BRCA-defect scores across HRD/BRCA-defect groups in the TCGA testing

1291 cohort. Each panel corresponds to a specific HRD/BRCA-defect signature, with the y-axis representing
1292 correlation with the respective centroid model. Each box represents refers to the samples within the
1293 respective group.

1294

1295

1296

1297

1298

1299

1300

1301

1302

1303

1304

1305

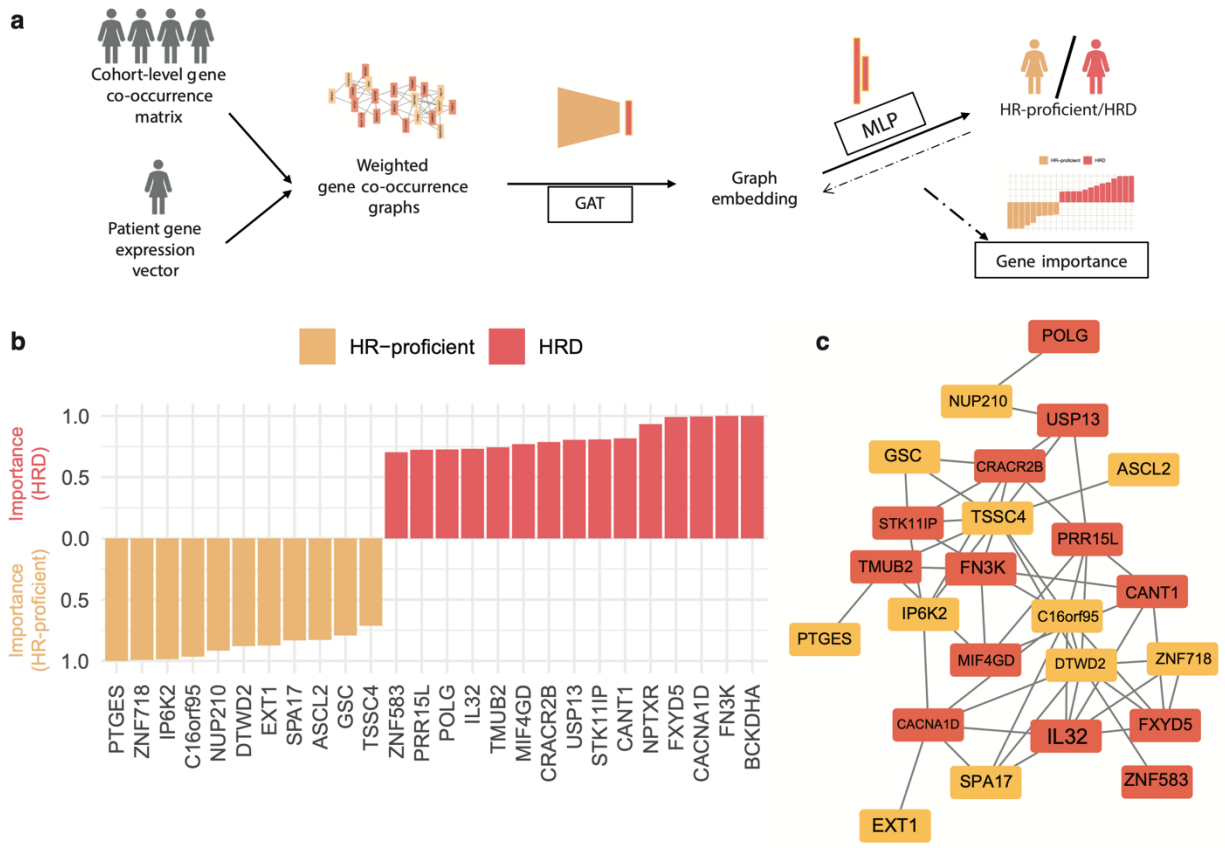

Figure 4. Graph analysis to determine transcriptional signature drivers. (a) Workflow for graph attention network analysis to classify HRD/HR-proficient TCGA-BRCA patients and determine gene importance. Weighted gene co-expression graphs are built from the gene expression profiles of the TCGA-BRCA cohort, while taking into account the patient-level gene expression for the 228 genes in the transcriptional HRD signature. A graph attention network (GAT) is then trained to distinguish HRD and HR-proficient samples using the weighted co-expression graphs as inputs. The output highlights part of the graphs with greater weight in the classification and generates an importance score for each genes. (b) The top-ranked 26 genes in the HRD versus HR-proficiency classification with importance scores greater than 0.7. The colour indicates the phenotype (HRD/HR-proficient) for which the gene is predictive. (c) The co-expression graph of 24 of the 26 highly ranked genes for classification. Genes are connected only if they are co-expressed in the cohort, and genes with no connections have been removed. The colour of the nodes depicts the associated phenotype as in (b).

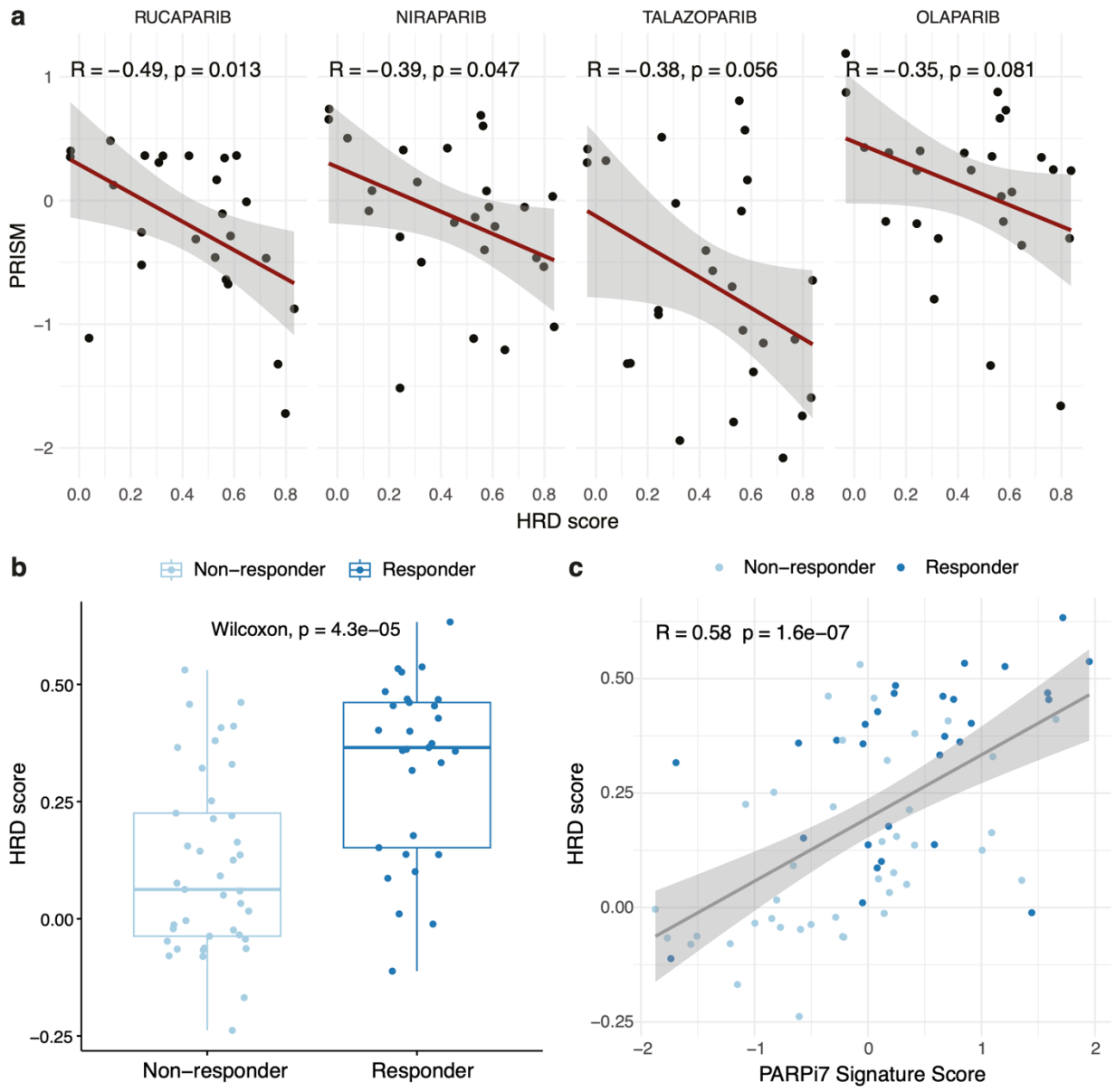

Figure 5. HRD transcriptional signature is linked with PARP inhibitor sensitivity in breast cancer cell lines and patients. (a) Correlation between the HRD transcriptional scores calculated using transcriptomes from breast cancer cell lines from CCLE and sensitivity to PARP inhibitors evaluated using the PRISM metric. (b) Comparison of HRD transcriptional scores between responders and non-responders to olaparib and durvalumab combination treatment in the I-SPY2 trial. (c) Correlation of HRD transcriptional scores against the PARPi7 signature score calculated by Pusztai et al.(38) in the I-SPY2 treatment arm patients. Responders are more frequently scoring high using our signature compared to PARPi7.

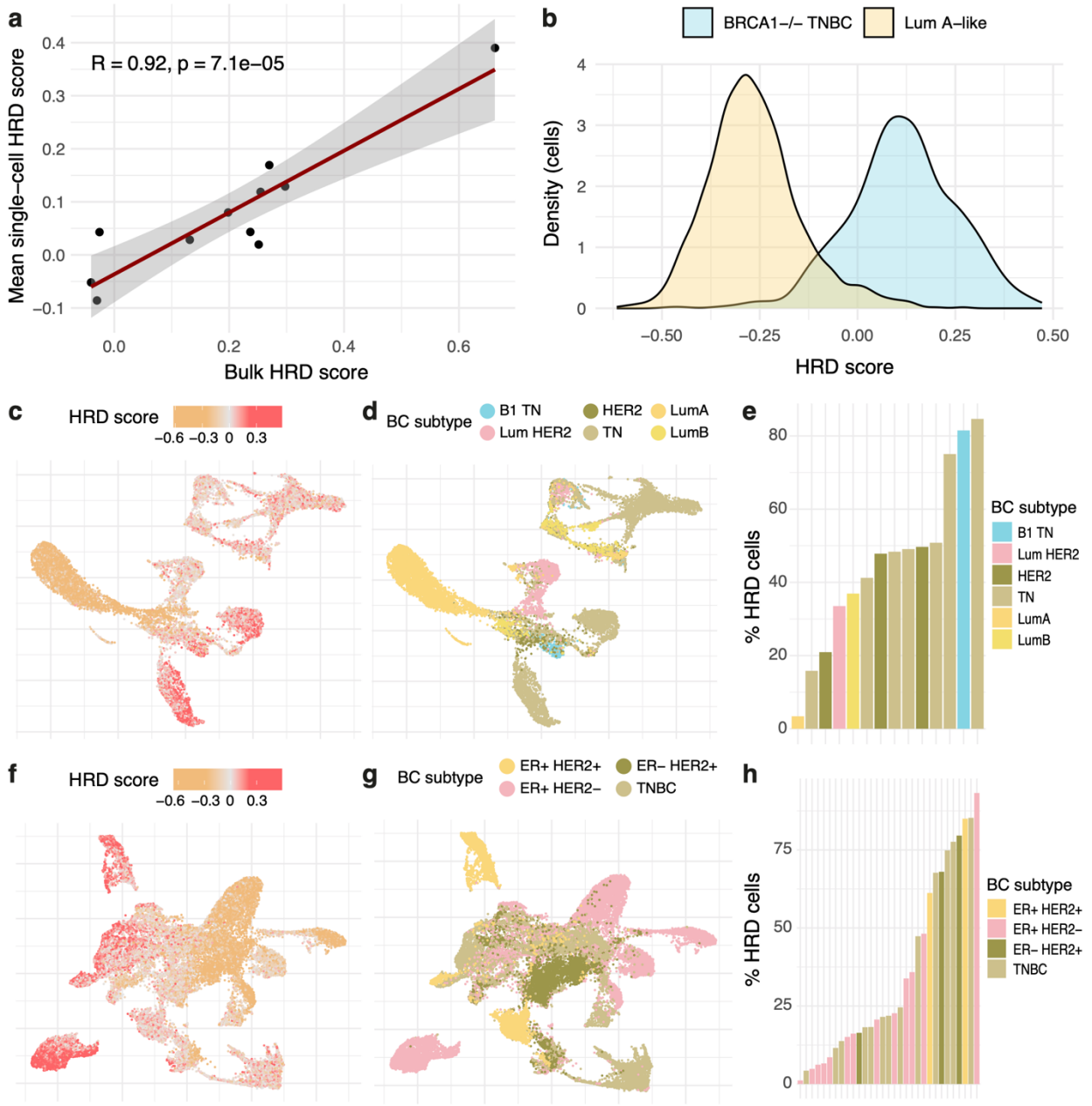

Figure 6. Transcriptional profiling of HRD in single cell sequenced breast cancer cells. (a) Correlation of mean HRD transcriptional score across individual cancer cells against matched bulk RNA sequencing from the Chung et al cohort(108). (b) Distribution of HRD scores across tumour cells from a Stage III BRCA1-defective TNBC sample (sc5rJUQ033) and a Stage II Luminal A sample (sc5rJUQ064) from the Qian et al cohort(71). (c-e) Profiling of HRD across tumour cells from the Qian et al cohort as demonstrated by UMAP coordinates labelled by c) HRD score and d) breast cancer subtype. (e) The proportion of cells within each sample with HRD scores greater than zero in the Qian et al cohort. The defined breast cancer subtypes include here are: 'B1 TN' = BRCA1-defective Triple negative, 'Lum HER2' = Luminal-HER2+, 'HER2' = HER2 positive, 'TN' = Triple negative, 'LumA' = Luminal A-like, 'LumB' = Luminal B-like (f-h) Profiling of HRD across tumour cells from the Bassez et al. cohort(85), similar to (c-e).

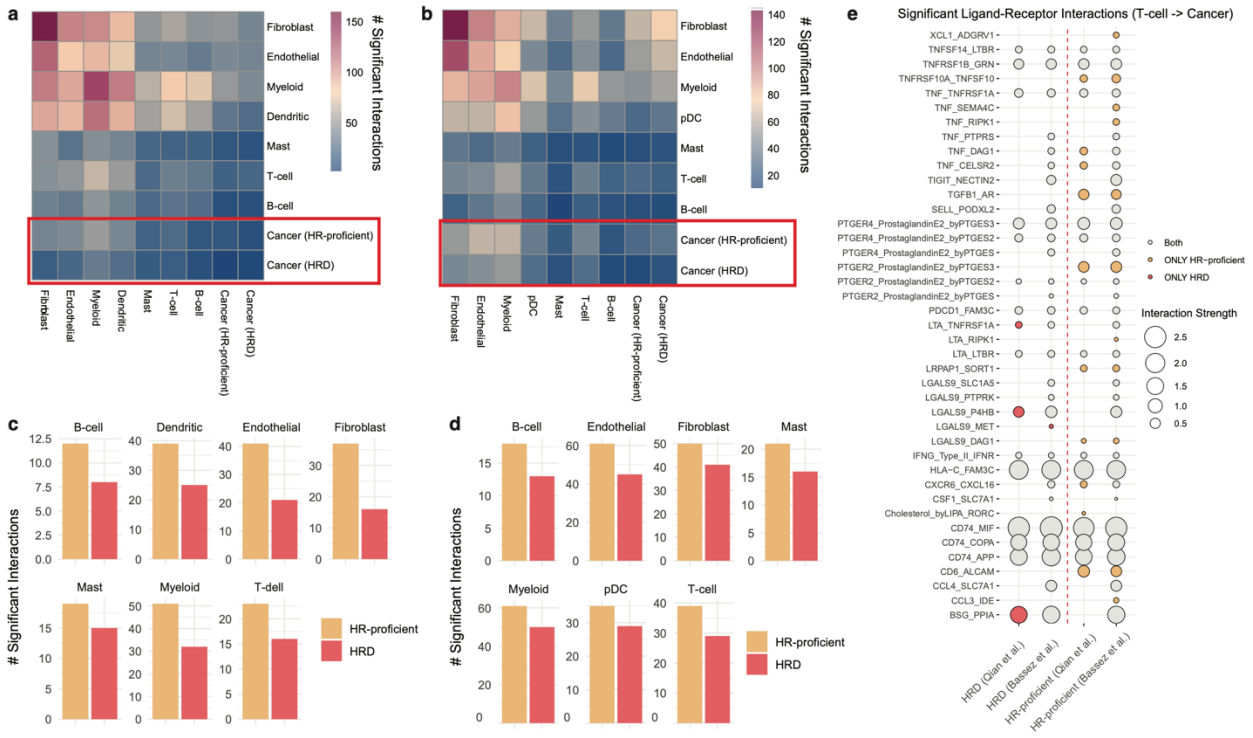

Figure 7. TME-cancer interactivity across HRD and HR-proficient cancer cells. (a-b) Number of significant ligand-receptor interactions established between cells in the (a) Qian et al.(71) and (b) Bassez et al.(85) cohorts according to CellphoneDB. Cancer cells are labelled as HRD if they have a positive HRD score, HR-proficient otherwise. The x-axis refers to cell types as sources, and the y-axis refers to cell types as targets. (c-d) The number of significant interactions between TME cell types as sources and cancer cells as targets, separated by HR status, across the (c) Qian et al. and (d) Bassez et al. cohorts. (e) Specific ligand-receptor interactions between T-cells and cancer cells, with cancer cells as the targets, across the Qian et al. and Bassez et al. cohorts. The red circles indicate interactions unique to HRD cells within a given cohort, and the yellow circles indicate interactions unique to HR-proficient cells within a given cohort. Grey circles represent common interactions.