

1 **Detecting HRD in whole-genome and whole-exome sequenced breast and ovarian**  
2 **cancers**

3 Ammal Abbasi<sup>1-5</sup>, Christopher D. Steele<sup>1-3</sup>, Erik N. Bergstrom<sup>1-3</sup>, Azhar Khandekar<sup>1-4</sup>,  
4 Akanksha Farswan<sup>6-7</sup>, Rana R. Mckay<sup>3</sup>, Nischalan Pillay<sup>6-7</sup>, and Ludmil B. Alexandrov<sup>1-5\*</sup>

5  
6 <sup>1</sup>Department of Cellular and Molecular Medicine, UC San Diego, La Jolla, CA, 92093, USA

7 <sup>2</sup>Department of Bioengineering, UC San Diego, La Jolla, CA, 92093, USA

8 <sup>3</sup>Moore's Cancer Center, UC San Diego, La Jolla, CA, 92037, USA

9 <sup>4</sup>Bioinformatics and Systems Biology Graduate Program, UC San Diego, La Jolla, CA,  
10 92093, USA

11 <sup>5</sup>Sanford Stem Cell Institute, University of California San Diego, La Jolla, CA 92037

12 <sup>6</sup>Research Department of Pathology, Cancer Institute, University College London,  
13 London, WC1E 6BT, UK

14 <sup>7</sup>Department of Cellular and Molecular Pathology, Royal National Orthopaedic Hospital  
15 NHS Trust, Stanmore, HA7 4LP, UK

16

17 \*Correspondence should be addressed to [L2alexandrov@health.ucsd.edu](mailto:L2alexandrov@health.ucsd.edu).

18

19 **ABSTRACT (147 words)**

20 Breast and ovarian cancers harboring homologous recombination deficiency (HRD) are  
21 sensitive to PARP inhibitors and platinum chemotherapy. Conventionally, detecting HRD  
22 involves screening for defects in *BRCA1*, *BRCA2*, and other relevant genes. Recent  
23 analyses have shown that HRD cancers exhibit characteristic mutational patterns due to  
24 the activities of HRD-associated mutational signatures. At least three machine learning  
25 tools exist for detecting HRD based on mutational patterns. Here, using sequencing data  
26 from 1,043 breast and 182 ovarian cancers, we trained Homologous Recombination  
27 Proficiency Profiler (HRProfiler), a machine learning method for detecting HRD using six  
28 mutational features. HRProfiler's performance is assessed against prior approaches  
29 using additional independent datasets of 417 breast and 115 ovarian cancers, including  
30 retrospective data from a clinical trial involving patients treated with PARP inhibitors. Our  
31 results demonstrate that HRProfiler is the only tool that robustly and consistently predicts  
32 clinical response from whole-exome sequenced breast and ovarian cancers.

33

34 **SIGNIFICANCE (48 words)**

35 HRProfiler is a novel machine learning approach that harnesses only six mutational  
36 features to detect clinically useful HRD from both whole-genome and whole-exome  
37 sequenced breast and ovarian cancers. Our results provide a practical way for detecting  
38 HRD and caution against using individual HRD-associated mutational signatures as  
39 clinical biomarkers.

## 40 INTRODUCTION

41 Repair of DNA double strand breaks by homologous recombination (HR) is an essential  
42 cellular mechanism for maintaining genomic stability and for preventing tumorigenesis  
43 (1). Prior studies have elucidated key genes in the HR pathway, including, *BRCA1*,  
44 *BRCA2*, *RAD51*, and *PALB2*, that commonly have pathogenic germline variants and/or  
45 somatic mutations in breast and ovarian cancers (1). Defects in HR genes can disable  
46 the HR repair pathway making cells vulnerable to double strand breaks and, thus, provide  
47 a treatment opportunity. Specifically, patients with cancers harboring defective HR repair  
48 are sensitive to both poly (ADP-ribose) polymerase inhibitors (PARPi) and to platinum  
49 chemotherapy (2,3).

50  
51 Conventional stratification of HR deficient (HRD) and HR proficient (HRP) cancers  
52 involves screening for canonical genomic markers, including pathogenic germline  
53 variants and somatic copy number alterations in HR genes (4-6). Previous experimental  
54 studies (7) and genomics analyses (8) have also revealed that HRD cells exhibit  
55 characteristic patterns of somatic mutations due to the activities of HRD-associated  
56 mutational processes. Currently, there are at least seven mutational signatures that have  
57 been putatively associated with and/or utilized to detect HRD: (i) single base substitution  
58 (SBS) signatures SBS3 and SBS8 both characterized by generally flat, yet distinct,  
59 profiles (9); (ii) genomic rearrangement signatures RS3 and RS5 reflecting non-clustered  
60 tandem duplications and deletion, respectively (10); (iii) small insertions and deletions  
61 (ID) signatures ID6 and ID8, predominately encompassing indels at microhomologies

62 (11); and (iv) copy number (CN) signature CN17 characterized by large tandem  
63 duplications (12).

64  
65 At least three machine learning approaches have also been developed to capture HR  
66 deficient cancers by examining the patterns of somatic mutations found in cancer  
67 genomes: HRDetect (13), CHORD (14), and SigMA (15). HRDetect uses signatures  
68 SBS3, SBS8, RS3, RS5, and indels at microhomologies corresponding to ID6 and ID8 to  
69 detect HRD in breast cancers (13). CHORD is an alternative pan-cancer HRD prediction  
70 tool that does not rely on mutational signatures, but it rather uses 29 mutational features  
71 directly observed in cancer genomes (14). CHORD is more computationally efficient and  
72 prior studies have shown that it has an almost identical performance to the one of  
73 HRDetect (13). However, both CHORD and HRDetect use HRD-specific patterns of  
74 genomic rearrangements that can be only reliably detected from whole-genome  
75 sequencing (WGS) data (13,14). By excluding genomic rearrangements, HRDetect can  
76 also be applied to whole-exome sequencing (WES) data, albeit, with significantly  
77 diminished performance (13). Conversely, CHORD's implementation does not allow  
78 utilizing WES cancers. In contrast to CHORD and HRDetect, SigMA was developed to  
79 exclusively detect HRD-associated signature SBS3 from whole-genome, whole-exome,  
80 and targeted gene panel sequencing data with SigMA's focus being on panel sequencing  
81 data (15). Nevertheless, to be applied to a sample, SigMA requires at least five somatic  
82 mutations within the examined cancer (15). Based on Memorial Sloan Kettering Cancer  
83 Center's Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT) data

84 (16), this limits SigMA's applicability to approximately 37% of breast and ovarian cancers  
85 profiled with MSK-IMPACT targeted gene panel.

86

87 In this manuscript, we perform retrospective analyses to evaluate the clinical utility of  
88 canonical gene-based biomarkers, HRD-associated mutational signatures, and machine  
89 learning approaches to detect treatment sensitive breast and ovarian cancers. While the  
90 presence of individual HRD-associated mutational signatures are generally ineffective in  
91 detecting clinical response, existing machine learning tools can capture treatment  
92 sensitivity in WGS cancers but not in WES cancers. To address this limitation, we  
93 developed Homologous Recombination Proficiency Profiler (HRProfiler), a machine  
94 learning method that harnesses only six mutational features for detecting clinically  
95 actionable HRD from both whole-genome and whole-exome sequenced breast and  
96 ovarian cancers. Our findings offer a pragmatic approach to detect HRD in WES cancers  
97 and underscore the importance of exercising caution when considering individual HRD-  
98 associated mutational signatures as clinical biomarkers.

99

## 100 RESULTS

### 101 *Feature engineering and model training of HRProfiler*

102 To determine the set of robust HRD-associated mutational patterns that can be detected  
103 using WGS and WES cancers, we identified significantly enriched mutation types specific  
104 to somatic SBSs (9), IDs (11), and CNs (12). In particular, using previously developed  
105 schemas (9,11,12), we compared the types of somatic mutations enriched in HRD or HRP  
106 cancers. Comparisons were performed for whole-genome sequenced breast cancers  
107 using a subset of the Sanger Institute's 560 breast cancer genomes cohort (10) (Sanger-  
108 WGS-Breast; **Fig. 1a**) as well as for whole-exome sequenced breast cancers using a  
109 subset of TCGA's breast cancer cohort (17) (TCGA-WES-Breast; **Fig. 1b**). As previously  
110 done (13,14,18) patients were classified as HRD based on a combination of their genomic  
111 instability and the presence of pathogenic germline variants, somatic mutations, or  
112 methylation of *BRCA1* or *BRCA2*. Feature engineering and the subsequent training of  
113 HRProfiler was performed only on the designated training datasets (**Supplementary Fig.**  
114 **S1**).

115

116 At the SBS resolution, we observed a striking enrichment of C:G>T:A single base  
117 substitutions at 5'-NpCpG-3' context (mutated base underlined; N reflects any base) in  
118 HRP samples (**Fig. 1a-b**). This suggests that a relatively large proportion of mutations in  
119 HRP samples are C:G>T:A transitions at CpG sites when compared to HRD samples.  
120 Conversely, HRD samples were enriched for C:G>G:C single base substitutions at 5'-  
121 NpCpT-3' context. At the indel resolution, we observed an enrichment of deletions  
122 spanning at least 5 base pairs (bp) with flanking microhomology sequences across HRD

123 samples (**Fig. 1a-b**). These mutations are known to arise from the erroneous activities of  
124 the microhomology-mediated end joining or the single strand annealing DNA repair  
125 pathways in the absence of a functional HR pathway (19). At the copy number resolution,  
126 Loss of Heterozygosity (LOH) events spanning 1 to 40Mb and heterozygous events  
127 spanning 10 to 40Mb with a Total Copy Number (TCN) state between 3 and 9 were  
128 enriched in HRD samples (**Fig. 1a-b**). In contrast, very large (>40Mb) heterozygous  
129 segments with TCN between 2 and 4 were enriched in HRP samples (**Fig. 1a-b**). This  
130 finding suggests that very large diploid segments or regions that have undergone  
131 genome-doubling are enriched in HRP samples, in line with the observation that HRP  
132 samples are genomically stable and harbor relatively low copy number aberrations (18).

133

134 Based on these observations, we combined the mutational channels (**Methods**) into six  
135 genomic features: (i) genomic segments with LOH and sizes between 1 and 40  
136 megabases (abbreviated as LOH:1-40Mb); (ii) deletions spanning at least 5bp at  
137 microhomologies (DEL.5.MH); (iii) heterozygous genomic segments with TCN between 3  
138 and 9 and sizes between 10 and 40 megabases (3-9:HET:10-40Mb); (iv) C:G>G:C  
139 substitutions at 5'-NpCpT-3' context (N[C>G]T); (v) C:G>T:A substitutions at 5'-NpCpG-  
140 3' context (N[C>T]G); and (vi) heterozygous genomic segments with TCN between 2 and  
141 4, and sizes above 40 megabases (2-4:HET:>40Mb). To evaluate if these genomic  
142 features are sufficient to distinguish HRD and HRP samples, we performed principal  
143 component analysis (PCA) using the training data. We observed a separation between  
144 HRD from HRP samples across the two principal components for both WGS (**Fig. 1c**) and  
145 WES (**Fig. 1d**) breast cancers.

146 Next, using the six genomic features, we trained a machine learning tool, HRProfiler,  
147 based on a linear kernel support vector machine. HRProfiler comprises WGS and WES  
148 models that were trained using 371 samples from the Sanger-WGS-Breast (13) and 672  
149 samples from the TCGA-WES-Breast (17) datasets respectively (**Supplementary Fig.**  
150 **S1**). Ten-fold cross validation was conducted to determine the feature weights for the two  
151 trained models. As expected, features with positive weights (*i.e.*, LOH:1-40Mb,  
152 DEL.5.MH, 3-9:HET:10-40Mb, and N[C>G]T) were enriched in HRD samples, whereas  
153 features with negative weights (*i.e.*, N[C>T]G and 2-4:HET:>40Mb) were enriched in HRP  
154 samples (**Fig. 1e**).

155

## 156 **Comparing HRD detection methods in WGS and WES breast cancers**

157 In principle, two distinct approaches have been utilized to evaluate the performance of  
158 methods for detecting HRD. In their original publications, CHORD and HRDetect have  
159 relied on concordance between their predictions and prior HRD genomic annotations  
160 (13,14). This concordance can be quantified by area under the receiver operating  
161 characteristic curve (AUC) with both CHORD and HRDetect reporting AUCs above 0.90  
162 for WGS cancers (13,14). However, this type of comparison requires a ground truth for  
163 HRD and HRP cancers which, in most cases, is not straightforward to derive. The second  
164 approach relies on comparing clinical endpoints for HRD and HRP predicted cancers in  
165 patients treated with either chemotherapy or PARPi. The advantage of this approach is  
166 that it could provide immediate clinical relevance. Unfortunately, such comparisons  
167 require the availability of well annotated clinico-genomics datasets which are currently



168 limited especially at the whole-genome resolution. Here, we utilize both approaches to  
169 put HRProfiler in the context of previously developed methods.

170

171 To evaluate the performance of HRProfiler, SigMA, HRDetect, and CHORD in the context  
172 of HRD genomic ground truth annotations, we applied the four tools to an independent  
173 set of 237 whole-genome sequenced triple negative breast cancers (TNBCs) from the  
174 Sweden Cancerome Analysis Network – Breast project (SCAN-B; ClinicalTrials.gov  
175 identifier NCT02306096) (20) as well as to 71 held-out TCGA breast cancers which have  
176 been profiled using both whole-genome and whole-exome sequencing. Additionally, we  
177 applied the tools to an independent external WES dataset of 109 MSK-IMPACT breast  
178 cancers (21). All tools exhibited good AUC performance when applied to the WGS  
179 cancers (**Fig. 2a-b; Supplementary Fig. S2a-b**) while HRProfiler outperformed  
180 HRDetect and SigMA for WES breast cancers (**Fig. 2c-d; Supplementary Fig. S2c-d**).  
181 CHORD could not be applied to WES data. Importantly, HRProfiler was the only tool with  
182 AUCs above 0.90 across all WES and WGS breast cancer datasets (**Fig. 2**).

183

184 To evaluate the potential clinical utility of HRProfiler, SigMA, HRDetect, and CHORD in  
185 serving as predictive biomarkers for adjuvant chemotherapy treated breast cancers, we  
186 applied the tools to a subset of 145 whole-genome sequenced chemotherapy-treated  
187 TNBCs with information for interval disease-free survival (20). Additionally, the 145  
188 TNBCs were down-sampled to whole-exomes (dWES) to further assess the ability of each  
189 tool to predict HRD robustly at both whole-genome and whole-exome resolutions. As

190 previously reported (20), when applied to WGS breast cancers, HRDetect was able to  
191 identify 99 HRD samples which exhibited better survival when compared to the 46 HRP  
192 samples after adjusting for grade and age at diagnosis (hazard ratio [HR]=0.42; p-  
193 value=0.020; **Fig. 3a**). However, the tool exhibited markedly worse sample stratification  
194 on the dWES data (HR=0.54; p-value=0.092) with 39 samples (26.9% of all examined  
195 TNBCs) being differently annotated when compared to the WGS data. CHORD's  
196 performance on WGS samples was very similar to that of HRDetect (**Supplementary**  
197 **Fig. S3**), however, the tool cannot be applied to the dWES data. Applying SigMA to the  
198 145 TNBCs did not result in a statistically significant separation for either the WGS breast  
199 cancers (p-value=0.068) or the dWES data (p-value=0.94; **Fig. 3b**). In contrast,  
200 HRProfiler was able to better stratify breast cancers from both WGS (HR=0.40; p-  
201 value=0.021) and dWES data (HR=0.38; p-value=0.02; **Fig. 3c**). Importantly, only 9  
202 samples (6.2% of all examined TNBCs) were differently annotated by HRProfiler when  
203 the tool was applied to WGS and dWES data (**Fig. 3c**). Lastly, partitioning the 145 TNBCs  
204 based on the presence of defects in *BRCA1/2* or the presence of HRD-associated  
205 signatures SBS3 or CN17 did not result in statistically significant separation  
206 (**Supplementary Fig. S4**). Nevertheless, stratifying the 145 TNBCs based on the  
207 presence of ID6 was able to separate the breast cancers, but captured 41 fewer HRD  
208 patients compared to HRProfiler (HR=0.48; p-value=0.04; **Supplementary Fig. S4**).

209

210 **Comparing HRD detection methods in WES ovarian cancers**

211 To determine if the breast cancer specific mutational features can be generalized to  
212 another HRD-associated cancer, we trained an ovarian-specific whole-exome model  
213 using 182 high-grade serous carcinoma from the TCGA-Ovarian-WES dataset (17)  
214 (**Supplementary Fig. S5a**). As done for breast cancer, ten-fold cross validation was  
215 conducted for HRProfiler to determine the feature weights for the trained whole-exome  
216 model. Similar features to the ones observed in breast cancer were enriched in HRD and  
217 HRP ovarian cancers (**Supplementary Fig. S5b**). To examine the performance of  
218 HRProfiler, SigMA, and HRDetect in the context of HRD genomic ground truth  
219 annotations for whole-exome sequenced ovarian cancer, we applied the three tools to 40  
220 held-out TCGA ovarian samples as well as to an independent set of 50 MSK-IMPACT  
221 whole-exome sequenced ovarian cancers (21) (**Supplementary Fig. S6a-b**). For both  
222 datasets, HRProfiler outperformed the other two approaches by consistently exhibiting  
223 AUCs above 0.90 (**Supplementary Fig. S6a-b**).

224  
225 To assess the clinical utility of HRProfiler, SigMA, and HRDetect to serve as predictors of  
226 clinical outcome in ovarian cancer, we examined the progression free survival for an  
227 independent set of 25 high-grade ovarian cancers from a phase Ib PARPi clinical trial of  
228 olaparib in combination with the PI3K inhibitor buparlisib (BKM120; ClinicalTrials.gov  
229 identifier NCT01623349) (22). HRProfiler's annotations were able to separate PARPi  
230 treated samples based on progression free survival (HR=0.25; p-value=0.037; **Fig. 4**)  
231 with HRDetect also performing relatively well on these data (HR=0.32; p-value=0.056;  
232 **Fig. 4b**). Moreover, partitioning the 25 PARPi-treated ovarian cancers based on the  
233 presence of any of the HRD-associated signatures SBS3, CN17, or ID6 did not lead to

234 differences in survival endpoints (**Supplementary Fig. S7**). Lastly, annotating samples  
235 as HRD and HRP based on defects in *BRCA1/2* genes provided separation in progression  
236 free survival for the 25 PARPi-treated ovarian cancers (**Supplementary Fig. S7**).

237

## 238 **DISCUSSION**

239 There is an increasing momentum in precision oncology towards more comprehensive  
240 genomic profiling to identify complex biomarkers like HRD as part of routine clinical care  
241 (23). With continuing advances in sequencing technologies and the corresponding  
242 exponential decrease in their cost, clinical whole-exome sequencing is becoming  
243 increasingly more prevalent (24-26). To harness the clinical utility of whole-exome  
244 sequencing for predicting HRD, we present a novel machine learning approach called  
245 HRProfiler that utilizes a minimal set of six genomic features to predict HRD across both  
246 whole-genome and whole-exome sequenced breast and ovarian cancers. Unlike existing  
247 methods that focus solely on mutation types enriched in HRD samples (13-15), HRProfiler  
248 incorporates small and large-scale mutational events enriched in both HRD and HRP  
249 cancers. HRProfiler also circumvents the need for genomic rearrangements and  
250 mutational signature extraction, which can be unreliable especially when using sparse  
251 datasets derived from whole-exome sequencing data (11).

252

253 HRProfiler demonstrated comparable performance to existing approaches when applied  
254 to whole-genome sequencing data and the tool surpassed other machine learning  
255 methods when applied to whole-exome sequenced cancers. The sub-optimal  
256 performance of HRDetect on whole-exome sequenced tumors is perhaps unsurprising  
257 given that HRDetect was developed for whole-genome sequenced breast cancers and  
258 the original publication noted a poor performance for whole-exome sequenced tumors  
259 (13). In contrast, despite its tailored design for whole-exome and targeted panel  
260 sequencing data, SigMA exhibited comparatively limited performance in our tests. Indeed,

261 SigMA is a machine learning surrogate for detecting HRD-associated signature SBS3 and  
262 our results show that SBS3 alone is not a reliable predictor of survival even when detected  
263 by other tools. Similarly, other HRD-associated signatures, such as CN17 and ID6, did  
264 not provide consistent clinical separation for breast or ovarian cancers. Overall, these  
265 results indicate that the presence of an individual HRD-associated signature in a cancer  
266 sample does not necessarily indicate a clinically significant or an actionable event.

267

268 HRProfiler's ability to separate HRD samples sensitive to treatment with PARP inhibitors  
269 from whole-exome sequencing data opens additional opportunities for broadening  
270 treatment options to a wider patient population. Given the non-tissue-specific nature of  
271 the HRD mutational footprint, our six mutational features can be refined in the future to  
272 predict HRD status in other HRD-associated cancers, including prostate and pancreatic  
273 cancers. Such an effort will ideally require large sets with well annotated clinico-genomics  
274 datasets for both cancer types, which, to the best of our knowledge, are currently not  
275 available.

276

277 Although we assessed HRProfiler's performance using independent datasets  
278 encompassing 417 breast and 115 ovarian cancers, along with retrospective data from  
279 two clinical trials, we recognize the constraints posed by the use of relatively small sample  
280 sizes for some of the reported survival analyses. Future large-scale, independent, and  
281 purposefully designed clinical trials will be necessary to validate HRProfiler's capacity to  
282 serve as a predictive and/or prognostic biomarker for routine clinical decision making.  
283 Notwithstanding, HRProfiler provides a crucial link in utilizing the molecular phenotypic

284 changes of impaired DNA repair mechanisms for detecting homologous recombination  
285 deficiency in whole-exome sequenced cancers. Moreover, the tool provides a robust and  
286 consistent approach that allows detecting whole-exome sequenced cancers that are  
287 sensitive to PARP inhibitors.

288

## 289 **METHODS**

### 290 **Data sources and pre-processing**

291 In this study, previously published datasets were used for all feature engineering, model  
292 development, and validation for both whole-genome sequenced (WGS) and whole-  
293 exome sequenced (WES) breast and ovarian cancers.

294 For breast cancer, we downloaded CaVEman mutations and ASCAT allele-specific copy  
295 number for 560 Sanger breast cancers (10) from: [ftp://ftp.sanger.ac.uk/pub/cancer/Nik-](ftp://ftp.sanger.ac.uk/pub/cancer/Nik-ZainalEtAl-560BreastGenomes/)  
296 [ZainalEtAl-560BreastGenomes/](ftp://ftp.sanger.ac.uk/pub/cancer/Nik-ZainalEtAl-560BreastGenomes/). Additional WGS breast cancer datasets used in this  
297 study included the 237 Triple Negative Breast (TNBC) samples from the SCAN-B clinical  
298 trial (20). CaVEman somatic mutations and ASCAT copy number for the 237 TNBC  
299 samples were downloaded from: <https://data.mendeley.com/datasets/2mn4ctdpxp/>. For  
300 the breast cancer WGS dataset from the Pan-Cancer Analysis of Whole Genomes  
301 project, consensus somatic mutations and copy number calls were downloaded from the  
302 International Cancer Genome Consortium's data portal:  
303 <https://dcc.icgc.org/releases/PCAWG>. For The Cancer Genome Atlas (TCGA) breast  
304 cancer WES dataset, the catalogues of somatic mutations and sequencing data were  
305 downloaded from the genomics data commons (<https://portal.gdc.cancer.gov/>) portal and  
306 allele-specific whole-exome copy number calls were derived using ASCAT:  
307 <https://github.com/VanLoo-lab/ascat>. For the WES MSK-IMPACT breast cancers, 109  
308 whole-exome sequenced breast cancers were downloaded from dbGaP (accession  
309 number: phs001783.v1.p1) and processed using an ensemble variant calling pipeline:  
310 <https://github.com/AlexandrovLab/EnsembleVariantCallingPipeline>



311 For ovarian cancer, the WES derived catalogues of somatic mutations and sequencing  
312 data from TCGA were downloaded from the genomics data commons portal, and allele-  
313 specific whole-exome copy number calls were derived using ASCAT. For the ovarian  
314 cancer WES MSK-IMPACT dataset, 50 whole-exome sequenced ovarian cancers were  
315 downloaded from dbGaP (accession number: phs001783.v1.p1) and processed using the  
316 same ensemble variant calling pipeline as the one utilized for breast cancer. Lastly, we  
317 downloaded the 25 PARPi treated high-grade ovarian cancers from dbGaP (accession  
318 number: phs003019) and processed these data using the ensemble variant calling  
319 pipeline.

320

### 321 **Feature engineering for predicting HRD**

322 As previously done (13,14,18), a sample with an HRD score of at least 42 for breast  
323 cancer (5) and 63 for ovarian cancer (27) or one harboring germline/somatic alterations  
324 in *BRCA1* or *BRCA2* was annotated as homologous recombination deficient (HRD) for all  
325 training purposes. All other samples were annotated as homologous recombination  
326 proficient (HRP). To identify significantly enriched features in HRD and HRP samples, we  
327 generated the average mutational profiles based on proportions across the 96  
328 substitution, 83 indel, and 48 copy number mutational contexts. To determine differences  
329 in channels at every resolution, we performed Fisher's exact tests to evaluate if there is  
330 any statistically significant difference in the average proportion of a given channel  
331 between HRD and HRP samples. Significant channels were identified for all types of  
332 mutational contexts if their absolute  $\log_2$  fold-change (FC) was greater than 0.75 for WGS  
333 samples and 0.25 for WES samples, and their  $-\log_{10}$ (FDR adjusted p-value) was greater

334 than 3. Similar workflow was adopted for both whole-genome and whole-exome samples  
335 and only channels significantly enriched across both WGS and WES were considered for  
336 the feature engineering process. At the single base resolution, A[C>T]G, C[C>T]G,  
337 G[C>T]G, and T[C>T]G channels were consistently enriched across HRP samples in both  
338 whole-genome and whole-exome datasets. Due to the overlapping/similar mutational  
339 context, these four channels were combined into a single feature termed N[C>T]G, where  
340 N represents any of the four nucleotide bases (A, C, T, or G). Similarly, A[C>G]T,  
341 C[C>G]T, G[C>G]T, and T[C>G]T were channels consistently enriched in HRD samples  
342 and were combined into a single feature N[C>G]T. At the indel resolution, 5:Del:M:1,  
343 5:Del:M:2, 5:Del:M:3, 5:Del:M:4, and 5:Del:M:5 were significantly enriched channels in  
344 HRD samples that represent varying lengths of microhomology sequences at relatively  
345 large deletion sites where the length of the deletion is at least 5 base pairs long. These  
346 indel channels were combined into a single feature: DEL.5.MH, where DEL.5 presents  
347 deletions of length at least 5 bp and MH represent microhomology sequences. At the  
348 copy number resolution, multiple significant Loss of Heterozygosity (LOH) events were  
349 identified. These events represented LOH segments of at least 1 Mb, where majority of  
350 the segment sizes ranged between 1 and 40Mb. These were combined into a single  
351 feature LOH:1-40Mb. A similar approach was applied to aggregate significant copy  
352 number channels for diploid/genome-doubled copy number segments into a single  
353 feature 2-4:HET:>40Mb that accounts for segments with a total copy number state  
354 between 2-4 and sizes of at least 40Mb. Lastly, significant copy number channels for  
355 amplification events were combined into a single feature: 3-9:HET:10-40Mb, where 3-9

356 represents the segments with a total copy number state of at least 3 and segment sizes  
357 between 10 to 40Mb.

### 358 **Training and comparing HRD detection methods in WGS cancers**

359 To train a model for predicting HRD at WGS resolution, we used samples from the 560  
360 Breast dataset. Only 371/560 samples that were labelled as evaluated in the HRDetect  
361 publication (13) were considered. The six features derived from the feature engineering  
362 step were extracted from the 371 samples and were normalized using StandardScaler in  
363 python's sklearn package. The training was based on 371 breast samples, comprising  
364 131 HRD and 240 HRP samples, and used a linear kernel support vector machine with  
365 L2 regularization. Next, 10-fold cross validation was conducted to tune for hyper-  
366 parameters and obtain feature weights from the model. To test the model's performance,  
367 we predicted HRD probabilities for 71 WGS TCGA breast samples that were sequenced  
368 at both whole-genome and whole-exome resolutions. Samples with an HRD probability  
369 at least 0.50 were considered as HRD. To validate the model on an external dataset, we  
370 predicted HRD probabilities for 237 Triple Negative Breast (TNBC) samples and  
371 evaluated its performance against the ground truth. The performance of the model was  
372 assessed using machine learning metrics such as sensitivity, precision, and F<sub>1</sub> score. To  
373 compare the performance of HRProfiler with other tools, HRD annotations were  
374 determined for the 237 TNBC samples using HRDetect, CHORD, and SigMA.

375

### 376 **Training and comparing HRD detection methods in WES cancers**

377 To train a breast cancer specific model for predicting HRD at WES resolution, we used  
378 samples from TCGA breast cancer dataset. Only 743 samples that had HRD annotations

379 were used for both training and testing. The six features derived from the feature  
380 engineering step were extracted as proportions, except for DEL.5.MH, which was  
381 extracted as absolute counts. All features were normalized using StandardScaler in  
382 python's sklearn package. The training was based on 672 breast samples that included  
383 156 HRD and 516 HRP samples. Next, 10-fold cross validation was conducted to tune  
384 for hyper-parameters and obtain feature weights from the model. The model's  
385 performance was tested on the held-out 71 breast samples that were previously  
386 sequenced at both whole-genome and whole-exome resolution. Samples with an HRD  
387 probability at least 0.50 were considered as HRD. To validate the model on an external  
388 dataset, we predicted HRD probabilities for 109 MSK-IMPACT breast cancer whole-  
389 exome sequenced samples and evaluated the model's performance against the ground  
390 truth. The performance of the model was assessed using conventional machine learning  
391 metrics such as sensitivity, precision, and  $F_1$  score. The WES model was also applied to  
392 the down-sampled 237 TNBC samples. The whole-exome features for the 237 TNBC  
393 samples were derived by down-sampling the ASCAT copy number calls to segments that  
394 spanned the exonic regions. The mutation and indel calls were down-sampled to whole-  
395 exome resolution using SigProfiler (28). To compare the performance of HRProfiler with  
396 other tools, HRD probabilities were also determined for SigMA and HRDetect.

397

398 To train an ovarian-specific model for predicting HRD at WES resolution, we used  
399 samples from the TCGA ovarian dataset. Only 228 samples that had HRD annotations  
400 were used for both training and testing. Analogous to training HRProfiler for WES breast  
401 cancers, the six features derived from the feature engineering step were extracted as

402 proportions, except for DEL.5.MH, which was extracted as absolute counts. All features  
403 were normalized using StandardScaler in the python sklearn package. The training was  
404 based on 182 ovarian cancers that comprised of 82 HRD and 100 HRP samples. Next,  
405 10-fold cross validation was conducted to tune for hyper-parameters and obtain feature  
406 weights from the model. The model's performance was tested on the 39 ovarian cancer  
407 that were sequenced at whole-exome resolution. Samples with an HRD probability at  
408 least 0.50 were considered as HRD. To validate the model on an external dataset, we  
409 predicted HRD probabilities for 50 MSK-IMPACT whole-exome sequenced ovarian  
410 cancers and evaluated the model's performance against the ground truth. The  
411 performance of the model was assessed using conventional machine learning metrics  
412 such as sensitivity, precision, and F<sub>1</sub> score. To compare the performance of HRProfiler  
413 with other tools, HRD annotations were determined for the same samples by HRDetect  
414 and SigMA using the default breast WGS and ovarian WES pre-trained models,  
415 respectively.

416

#### 417 **Deriving HRD status based on HRD-associated signatures, genes, and tools**

418 Germline and somatic mutations for *BRCA1* and *BRCA2* and, when available, gene  
419 expression and promoter methylation changes in *BRCA1* and *BRCA2* were incorporated  
420 for the *BRCA1/2* annotations. Specifically, for TCGA breast cancers, the *BRCA1/2*  
421 annotations were derived from Polak *et al.* (29). Conversely, for TCGA ovarian cancers,  
422 these annotations were derived from Steele *et al.* (12). For all other datasets, *BRCA1/2*  
423 annotations were derived from their respective publications.

424

425 SigProfilerAssignment (v0.1.2) was used to determine the presence of HRD-associated  
426 signatures SBS3, ID6, and CN17 (30) using the Catalogue Of Somatic Mutations In  
427 Cancer (COSMICv3.4) reference signatures. A sample was classified as HRD positive  
428 for a given HRD signature, if it had at least one mutational event attributed to that  
429 signature.

430

431 HRDetect was run using the Signature.tools.lib (v2.3.0) package in R, available at  
432 <https://github.com/Nik-Zainal-Group/signature.tools.lib>. The default HRD probability  
433 threshold of 0.70 was employed for predicting HRD status for WGS samples. To execute  
434 HRDetect on WES data, we utilized the pre-trained WGS model for prediction. The  
435 rearrangement signatures RS3 and RS5, which cannot be derived from WES data, were  
436 set to zero, and the default probability threshold of 0.70 was applied for classifying whole-  
437 exome sequenced cancers as HRD.

438

439 CHORD was run using the extractSigsChord function installed from GitHub:  
440 <https://github.com/UMCUGenetics/CHORD/>. It was executed using default settings, and  
441 a probability threshold of 0.50 was applied for classifying samples as HRD.

442

443 SigMA (v2.0) was downloaded from GitHub:  
444 <https://github.com/parklab/SigMA/archive/refs/tags/2.0.tar.gz> and it was run using the run  
445 function for signature 3 (also known as SBS3) prediction. For WGS breast datasets, we  
446 used the following parameters when running SigMA: data='wgs', do\_assign=T,  
447 do\_mva=T, tumor\_type='breast', and catalog\_name='cosmic\_v3p2\_inhouse', and we

448 utilized SigMA strict predictions (pass\_mva\_strict) for our analysis. When running SigMA  
449 on WES datasets, we followed the same procedure as for WGS datasets, except for the  
450 data and tumor\_type parameters. For predicting signature 3 status for TCGA datasets,  
451 the data parameter was set to 'tcga\_mc3', otherwise, it was set to 'seqcap' for all other  
452 WES and down-sampled WES datasets. The tumor\_type parameter was set to 'breast'  
453 for breast and 'ovary' for ovarian whole-exome sequencing data.

454

### 455 **Survival analysis**

456 The survival analysis was conducted using the KaplanMeierFitter and CoxPHFitter  
457 function from the lifelines package in python (31). Interval disease free survival was used  
458 to evaluate patients treated with chemotherapy from the 237 TNBC dataset. Progression  
459 free survival endpoint was used to evaluate the survival trends for 25 high-grade ovarian  
460 cancer patients treated with PARP inhibitor. P-values and hazard ratios listed in the  
461 Kaplan Meier plots are based on the p-values derived from the Cox proportional hazards  
462 (coxph) model adjusted by dichotomized age of diagnosis (below and above 50 years  
463 old) as well as tumor stage or grade.

464

### 465 **Statistics**

466 All statistical analysis were conducted in python using the scikit-learn package. All p-  
467 values were corrected for multiple hypothesis testing using Benjamini-Hochberg  
468 procedure, where applicable.

469

470

471 **Availability of data and materials**

472 HRProfiler is an open-source tool, and it is freely available for academic use as a python  
473 package at <https://github.com/AlexandrovLab/HRProfiler>. The pre-trained models for  
474 whole-genome and whole-exome sequenced breast and ovarian cancers are provided as  
475 part of the tool.

476

477 **Competing interests**

478 LBA is a co-founder, CSO, scientific advisory member, and consultant for io9, has equity  
479 and receives income. The terms of this arrangement have been reviewed and approved  
480 by the University of California, San Diego in accordance with its conflict of interest  
481 policies. LBA's spouse is an employee of Biotheranostics. ENB is a consultant for io9,  
482 has equity, and receives income. AA and LBA declare U.S. provisional patent application  
483 filed with UCSD with serial numbers 63/366,392 for detecting homologous recombination  
484 deficiency from genomics data. ENB and LBA declare U.S. provisional patent application  
485 filed with UCSD with serial numbers 63/269,033 for artificial intelligence architecture for  
486 predicting cancer biomarkers, including homologous recombination deficiency. LBA also  
487 declares U.S. provisional applications with serial numbers: 63/289,601; 63/483,237;  
488 63/412,835; and 63/492,348. All other authors declare that they have no competing  
489 interests.

490

491 **Funding and acknowledgements**

492 This work was supported by the US National Institute of Health grants R01ES030993-  
493 01A1, U01DE033345, R01ES032547-01, and R01CA269919-01 to LBA as well as LBA's



494 Packard Fellowship for Science and Engineering and Cancer Research UK Grand  
495 Challenge Award C98/A24032. The research in this grant was also supported by a  
496 Curebound Targeted grant to LBA and RRM. The computational analyses reported in this  
497 manuscript have utilized the Triton Shared Computing Cluster at the San Diego  
498 Supercomputer Center of UC San Diego. The funders had no roles in study design, data  
499 collection and analysis, decision to publish, or preparation of the manuscript.

500

### 501 **Authors' contributions**

502 AA and LBA designed the overall study. AA performed all analyses with help from CDS,  
503 ENB, AK, AF, RRM, NP, and LBA. CDS and AK assisted in the copy number analysis  
504 and the feature engineering process. AF contributed to testing the tool's functionality.  
505 ENB, RRM, NP, and LBA assisted in the interpretation and analysis of the survival and  
506 clinical associations. AA and LBA wrote the manuscript with help and input from all other  
507 authors. All authors read and approved the final manuscript.

508

## 509 REFERENCES

- 510 1. Konstantinopoulos PA, Ceccaldi R, Shapiro GI, D'Andrea AD. Homologous  
511 Recombination Deficiency: Exploiting the Fundamental Vulnerability of Ovarian  
512 Cancer. *Cancer Discov* **2015**;5(11):1137-54 doi 10.1158/2159-8290.CD-15-0714.
- 513 2. Moore K, Colombo N, Scambia G, Kim BG, Oaknin A, Friedlander M, *et al.*  
514 Maintenance Olaparib in Patients with Newly Diagnosed Advanced Ovarian  
515 Cancer. *N Engl J Med* **2018**;379(26):2495-505 doi 10.1056/NEJMoa1810858.
- 516 3. Tutt A, Tovey H, Cheang MCU, Kernaghan S, Kilburn L, Gazinska P, *et al.*  
517 Carboplatin in BRCA1/2-mutated and triple-negative breast cancer BRCAness  
518 subgroups: the TNT Trial. *Nat Med* **2018**;24(5):628-37 doi 10.1038/s41591-018-  
519 0009-7.
- 520 4. Birkbak NJ, Wang ZC, Kim JY, Eklund AC, Li Q, Tian R, *et al.* Telomeric allelic  
521 imbalance indicates defective DNA repair and sensitivity to DNA-damaging  
522 agents. *Cancer Discov* **2012**;2(4):366-75 doi 10.1158/2159-8290.CD-11-0206.
- 523 5. Telli ML, Timms KM, Reid J, Hennessy B, Mills GB, Jensen KC, *et al.* Homologous  
524 Recombination Deficiency (HRD) Score Predicts Response to Platinum-  
525 Containing Neoadjuvant Chemotherapy in Patients with Triple-Negative Breast  
526 Cancer. *Clin Cancer Res* **2016**;22(15):3764-73 doi 10.1158/1078-0432.CCR-15-  
527 2477.
- 528 6. Abkevich V, Timms KM, Hennessy BT, Potter J, Carey MS, Meyer LA, *et al.*  
529 Patterns of genomic loss of heterozygosity predict homologous recombination  
530 repair defects in epithelial ovarian cancer. *Br J Cancer* **2012**;107(10):1776-82 doi  
531 10.1038/bjc.2012.451.
- 532 7. Zamborszky J, Szikriszt B, Gervai JZ, Pipek O, Poti A, Krzystanek M, *et al.* Loss  
533 of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis  
534 and has distinct effects on genomic deletions. *Oncogene* **2017**;36(35):5085-6 doi  
535 10.1038/onc.2017.213.
- 536 8. Petljak M, Alexandrov LB, Brammell JS, Price S, Wedge DC, Grossmann S, *et al.*  
537 Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals  
538 Episodic APOBEC Mutagenesis. *Cell* **2019**;176(6):1282-94 e20 doi  
539 10.1016/j.cell.2019.02.012.
- 540 9. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, *et al.*  
541 Signatures of mutational processes in human cancer. *Nature* **2013**;500(7463):415-  
542 21 doi 10.1038/nature12477.
- 543 10. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, *et al.*  
544 Landscape of somatic mutations in 560 breast cancer whole-genome sequences.  
545 *Nature* **2016**;534(7605):47-54 doi 10.1038/nature17676.
- 546 11. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, *et al.* The  
547 repertoire of mutational signatures in human cancer. *Nature* **2020**;578(7793):94-  
548 101 doi 10.1038/s41586-020-1943-3.
- 549 12. Steele CD, Abbasi A, Islam SMA, Bowes AL, Khandekar A, Haase K, *et al.*  
550 Signatures of copy number alterations in human cancer. *Nature*  
551 **2022**;606(7916):984-91 doi 10.1038/s41586-022-04738-6.

- 552 13. Davies H, Glodzik D, Morganella S, Yates LR, Staaf J, Zou X, *et al.* HRDetect is a  
553 predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat*  
554 *Med* **2017**;23(4):517-25 doi 10.1038/nm.4292.
- 555 14. Nguyen L, J WMM, Van Hoeck A, Cuppen E. Pan-cancer landscape of  
556 homologous recombination deficiency. *Nat Commun* **2020**;11(1):5584 doi  
557 10.1038/s41467-020-19406-4.
- 558 15. Gulhan DC, Lee JJ, Melloni GEM, Cortes-Ciriano I, Park PJ. Detecting the  
559 mutational signature of homologous recombination deficiency in clinical samples.  
560 *Nat Genet* **2019**;51(5):912-9 doi 10.1038/s41588-019-0390-2.
- 561 16. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, *et al.* Mutational  
562 landscape of metastatic cancer revealed from prospective clinical sequencing of  
563 10,000 patients. *Nat Med* **2017**;23(6):703-13 doi 10.1038/nm.4333.
- 564 17. Gao GF, Parker JS, Reynolds SM, Silva TC, Wang LB, Zhou W, *et al.* Before and  
565 After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons'  
566 Data. *Cell Syst* **2019**;9(1):24-34 e10 doi 10.1016/j.cels.2019.06.006.
- 567 18. Marquard AM, Eklund AC, Joshi T, Krzystanek M, Favero F, Wang ZC, *et al.* Pan-  
568 cancer analysis of genomic scar signatures associated with homologous  
569 recombination deficiency suggests novel indications for existing cancer drugs.  
570 *Biomark Res* **2015**;3:9 doi 10.1186/s40364-015-0033-4.
- 571 19. Pettitt SJ, Frankum JR, Punta M, Lise S, Alexander J, Chen Y, *et al.* Clinical  
572 *brca1/2* reversion analysis identifies hotspot mutations and predicted neoantigens  
573 associated with therapy resistance. *Cancer Discovery* **2020**;10(10):1475-88 doi  
574 10.1158/2159-8290.CD-19-1485.
- 575 20. Staaf J, Glodzik D, Bosch A, Vallon-Christersson J, Reutersward C, Hakkinen J,  
576 *et al.* Whole-genome sequencing of triple-negative breast cancers in a population-  
577 based clinical study. *Nat Med* **2019**;25(10):1526-33 doi 10.1038/s41591-019-  
578 0582-4.
- 579 21. Jonsson P, Bandlamudi C, Cheng ML, Srinivasan P, Chavan SS, Friedman ND, *et*  
580 *al.* Tumour lineage shapes BRCA-mediated phenotypes. *Nature*  
581 **2019**;571(7766):576-9 doi 10.1038/s41586-019-1382-1.
- 582 22. Batalini F, Gulhan DC, Mao V, Tran A, Polak M, Xiong N, *et al.* Mutational  
583 Signature 3 Detected from Clinical Panel Sequencing is Associated with  
584 Responses to Olaparib in Breast and Ovarian Cancers. *Clin Cancer Res*  
585 **2022**;28(21):4714-23 doi 10.1158/1078-0432.CCR-22-0749.
- 586 23. Menzel M, Ossowski S, Kral S, Metzger P, Horak P, Marienfeld R, *et al.* Multicentric  
587 pilot study to standardize clinical whole exome sequencing (WES) for cancer  
588 patients. *NPJ Precis Oncol* **2023**;7(1):106 doi 10.1038/s41698-023-00457-x.
- 589 24. Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marlow S, *et al.* Whole-  
590 exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded  
591 tumor samples to guide precision cancer medicine. *Nat Med* **2014**;20(6):682-8 doi  
592 10.1038/nm.3559.
- 593 25. Horak P, Heining C, Kreutzfeldt S, Hutter B, Mock A, Hullein J, *et al.*  
594 Comprehensive Genomic and Transcriptomic Analysis for Guiding Therapeutic  
595 Decisions in Patients with Rare Cancers. *Cancer Discov* **2021**;11(11):2780-95 doi  
596 10.1158/2159-8290.CD-21-0126.

- 597 26. Niguidula N, Alamillo C, Shahmirzadi Mowlavi L, Powis Z, Cohen JS, Farwell  
598 Hagman KD. Clinical whole-exome sequencing results impact medical  
599 management. *Mol Genet Genomic Med* **2018**;6(6):1068-78 doi  
600 10.1002/mgg3.484.
- 601 27. Takaya H, Nakai H, Takamatsu S, Mandai M, Matsumura N. Homologous  
602 recombination deficiency status-based classification of high-grade serous ovarian  
603 carcinoma. *Sci Rep* **2020**;10(1):2757 doi 10.1038/s41598-020-59671-3.
- 604 28. Bergstrom EN, Huang MN, Mahto U, Barnes M, Stratton MR, Rozen SG, *et al.*  
605 SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small  
606 mutational events. *BMC Genomics* **2019**;20(1):685 doi 10.1186/s12864-019-6041-  
607 2.
- 608 29. Polak P, Kim J, Braunstein LZ, Karlic R, Haradhavala NJ, Tiao G, *et al.* A  
609 mutational signature reveals alterations underlying deficient homologous  
610 recombination repair in breast cancer. *Nat Genet* **2017**;49(10):1476-86 doi  
611 10.1038/ng.3934.
- 612 30. Diaz-Gay M, Vangara R, Barnes M, Wang X, Islam SMA, Vermes I, *et al.* Assigning  
613 mutational signatures to individual samples and individual somatic mutations with  
614 SigProfilerAssignment. *Bioinformatics* **2023**;39(12) doi  
615 10.1093/bioinformatics/btad756.
- 616 31. Davidson-Pilon C. lifelines: survival analysis in Python. *Journal of Open Source*  
617 *Software* **2019**;4(40) doi 10.21105/joss.01317.  
618

619 **FIGURE LEGENDS**

620 **Figure 1: Feature engineering to identify significantly enriched somatic mutational**  
621 **features across HRD and HRP breast cancers. (a-b)** Volcano plots with  $\log_2$  fold  
622 change (FC) enrichments across the average proportions of somatic mutations for 96  
623 substitution, 83 indel, and 48 copy number mutational channels between homologous  
624 recombination deficient (HRD) and homologous recombination proficient (HRP) cancers  
625 for 371 Sanger-WGS-Breast (a) and 672 TCGA-WES-Breast samples (b). Channels with  
626 an absolute FC greater than 0.75 for WGS and 0.25 for WES, and a  $-\log_{10}$  FDR adjusted  
627 p-value greater than 3 are colored. Channels colored in red are enriched in HRD samples,  
628 while channels highlighted in blue are enriched in HRP samples. **(c-d)** Principal  
629 component (PC) analysis highlights the relevance of the features derived from the  
630 significant channels in (a-b) by separating HRD from HRP samples across the 371  
631 Sanger-WGS-Breast (c) and 672 TCGA-WES-Breast cohorts (d). **(e)** The average 10-fold  
632 cross validation weights of the six features derived from the WGS and WES breast  
633 training datasets using a linear-kernel support vector machine. Positive weights reflect  
634 features predictive for HRD samples, while negative weights correspond to features  
635 predictive for HRP samples.

636  
637 **Figure 2: Performance of HRD tools on external validation datasets using HRD**  
638 **genomic ground truth annotations.** Receiver operating characteristic curves (ROCs)  
639 were derived for HRProfiler, SigMA, HRDetect, and CHORD. Areas under the ROCs  
640 (AUCs) were calculated for each tool and shown in the legends of the respective panels.  
641 **(a)** ROCs for 237 whole-genome sequenced (WGS) triple negative breast cancers. **(b)**

642 ROCs for 71 WGS TCGA breast cancers. **(c)** ROCs for 71 whole-exome sequenced  
643 (WES) breast cancers. **(d)** ROCs for 109 WES MSK-IMPACT breast cancers. No ROCs  
644 are shown for CHORD in panels *(c)* and *(d)* as the tool cannot be applied to WES data.  
645 In all plots, the x-axes reflect the false positive rates while the y-axes correspond to the  
646 true positive rates. Precision and recall curves for the same samples are provided in  
647 **Supplementary Figure S2.**

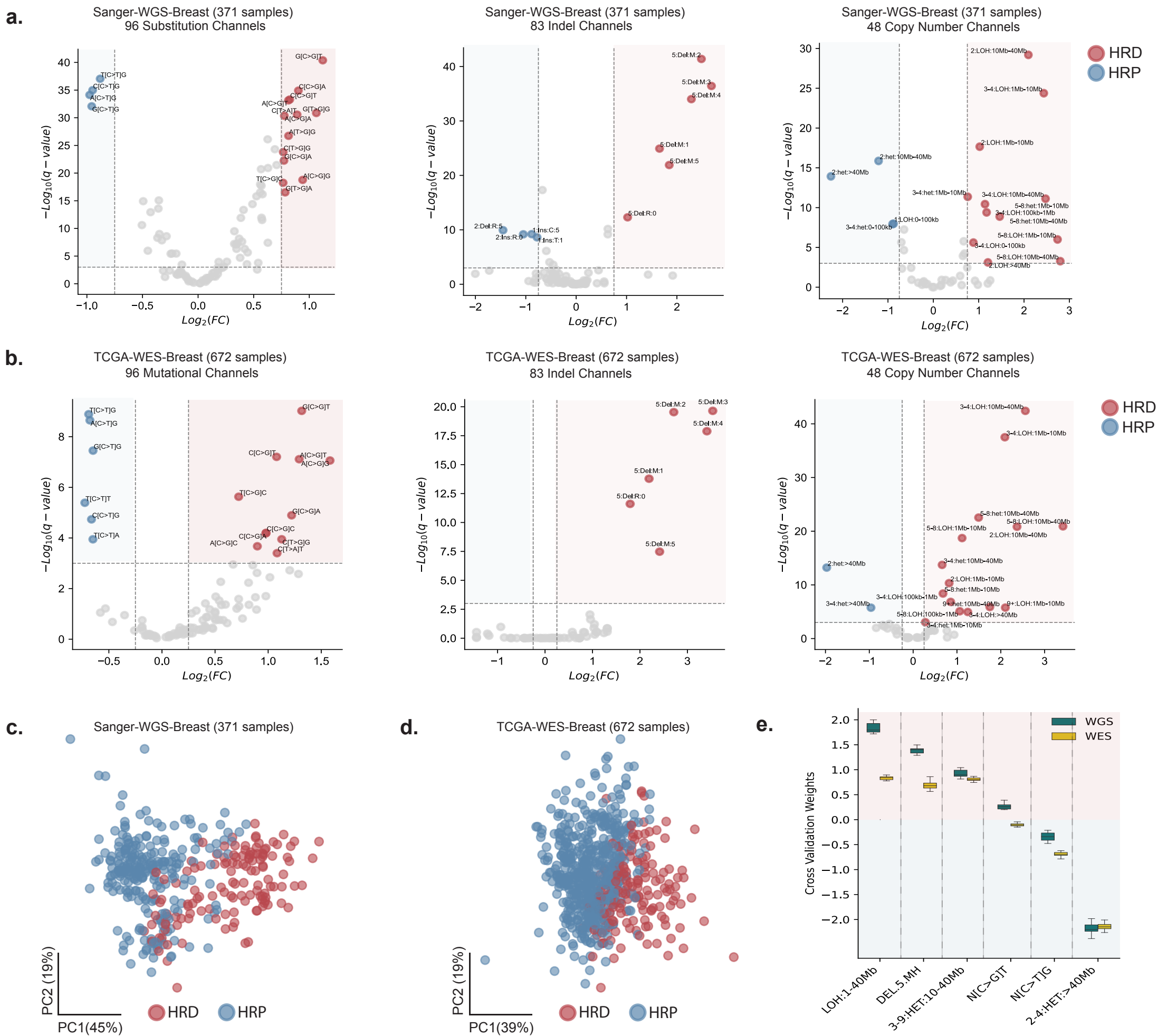
648  
649 **Figure 3: Predicting survival in breast cancers treated with chemotherapy by HRD**  
650 **tools.** Kaplan-Meier curves and confusion matrices for samples predicted as HRD and  
651 HRP by **(a)** HRDetect, **(b)** SigMA, and **(c)** HRProfiler in 145 chemotherapy-treated triple  
652 negative breast cancers. In each panel, the left plot reflects the Kaplan-Meier curves for  
653 whole-genome sequenced breast cancers (WGS). The middle plot corresponds to the  
654 Kaplan-Meier curves for the same samples when down-sampled to whole-exomes  
655 (dWESs). The right plot contains a confusion matrix that provides a comparison of each  
656 tool's HRD annotations from WGS and dWES data. The y-axes on all Kaplan-Meier  
657 curves reflect Interval Disease Free Survival (IDFS), and the x-axes correspond to time  
658 measured in years. Listed p-values and hazard ratios (HRs) are based on a Cox  
659 proportional hazards model after adjusting for age at diagnosis and tumor grade. 95%  
660 confidence intervals are provided for all HRs within the Kaplan-Meier plots. The  
661 performance of CHORD on WGS data, which was almost identical to the one of  
662 HRDetect, can be found in **Supplementary Figure S3**. Comparisons of the clinical utility  
663 of BRCA1/2 defects and HRD-associated signatures SBS3, CN17, and ID6 for the same  
664 patients are provided in **Supplementary Figure S4**.

665 **Figure 4: Predicting survival in ovarian cancers treated PARP inhibitor by HRD**  
666 **tools.** Kaplan-Meier curves for progression free survival (PFS) across 25 PARP inhibitor  
667 treated patients with high-grade serous ovarian cancer. Patients are annotated as HRD  
668 or HRP based on the predictions from HRProfiler (*left panel*), SigMA (*middle panel*), and  
669 HRDetect (*right panel*). Listed p-values and hazard ratios (HRs) are based on a Cox  
670 proportional hazards model after adjusting for age at diagnosis and tumor stage. 95%  
671 confidence intervals are provided for all HRs within the Kaplan-Meier plots. Comparisons  
672 of the clinical utility of BRCA1/2 defects and HRD-associated signatures SBS3, CN17,  
673 and ID6 for the same patients are provided in **Supplementary Figure S7**.

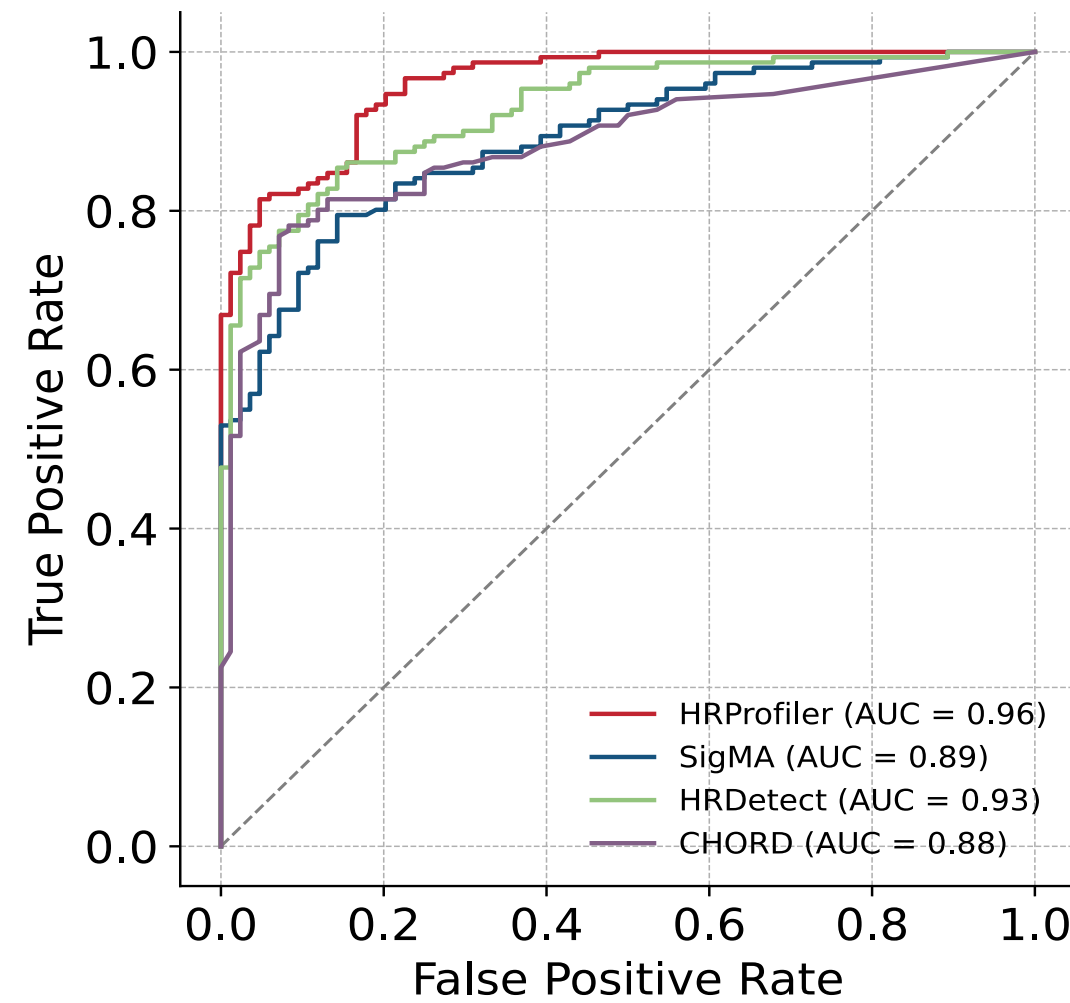
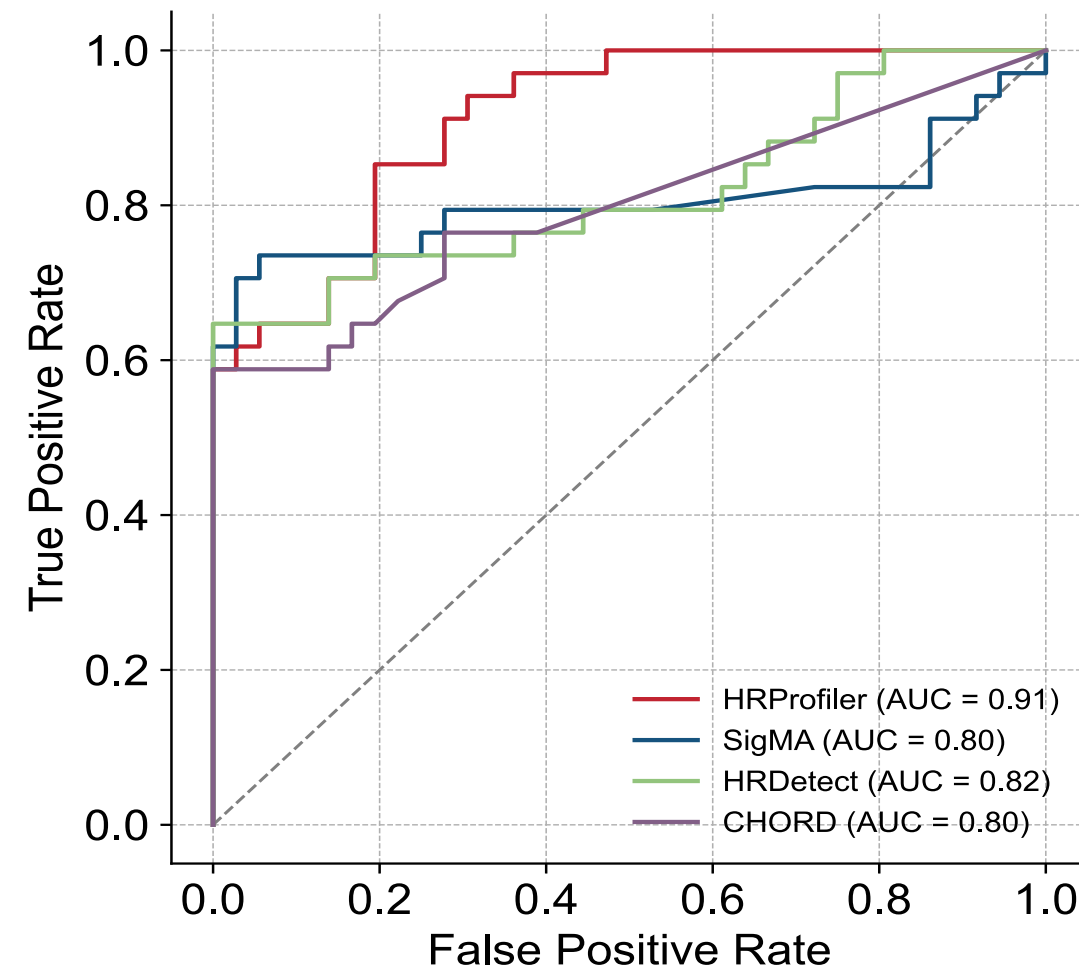
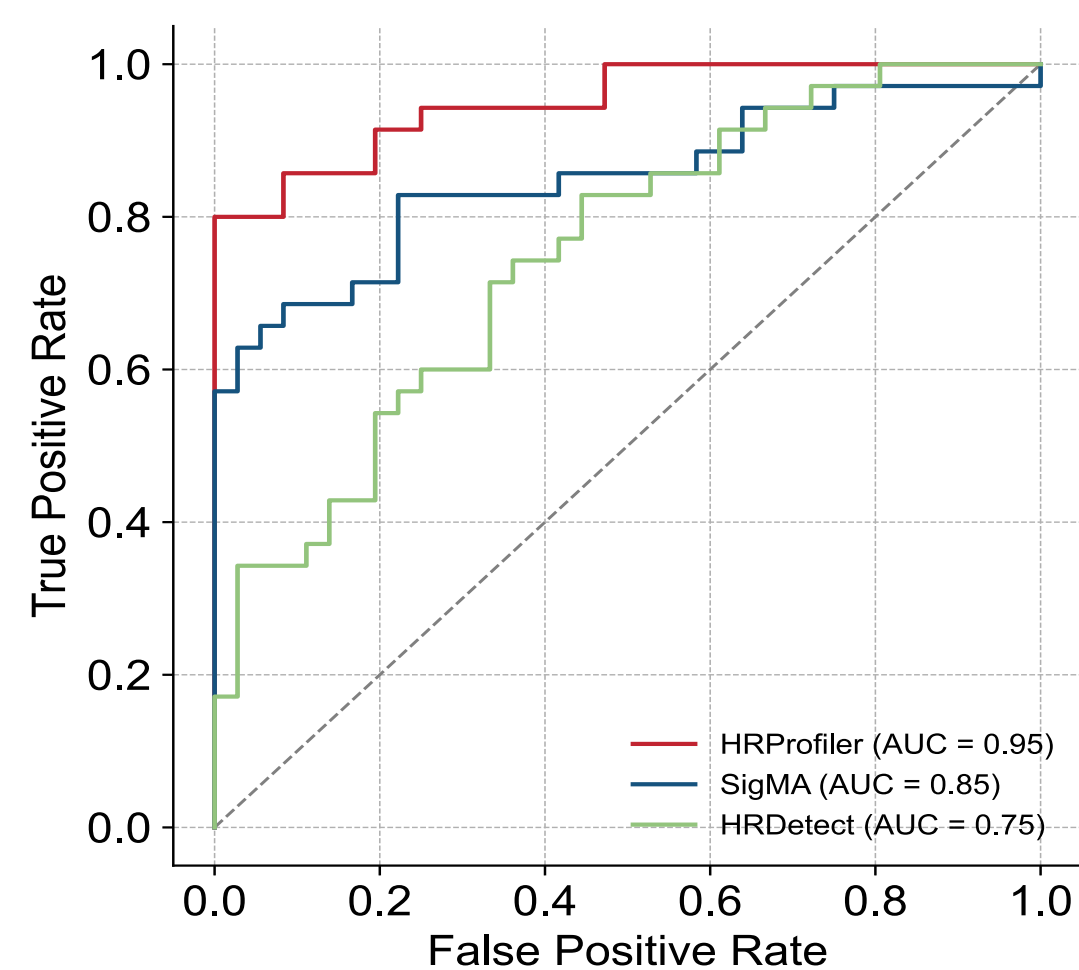
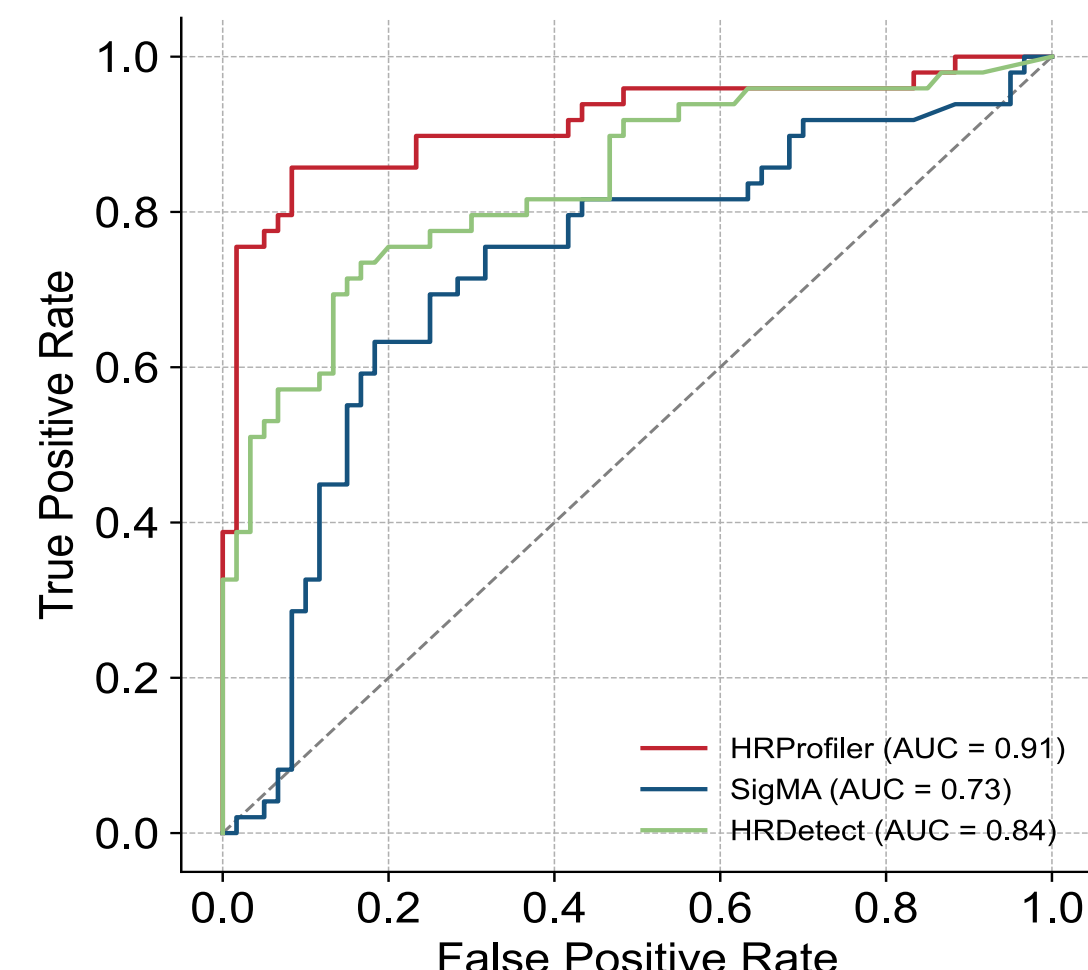


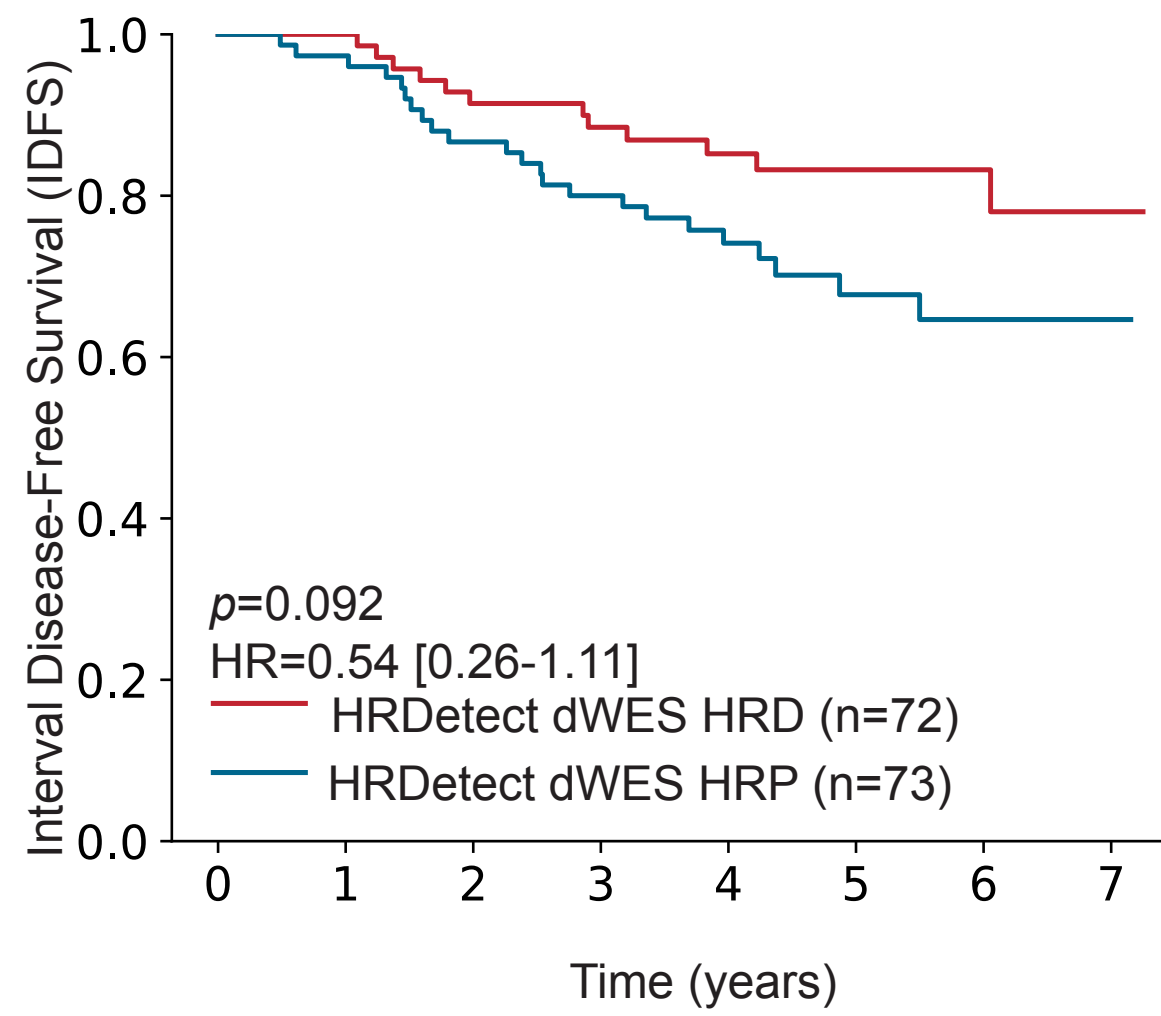
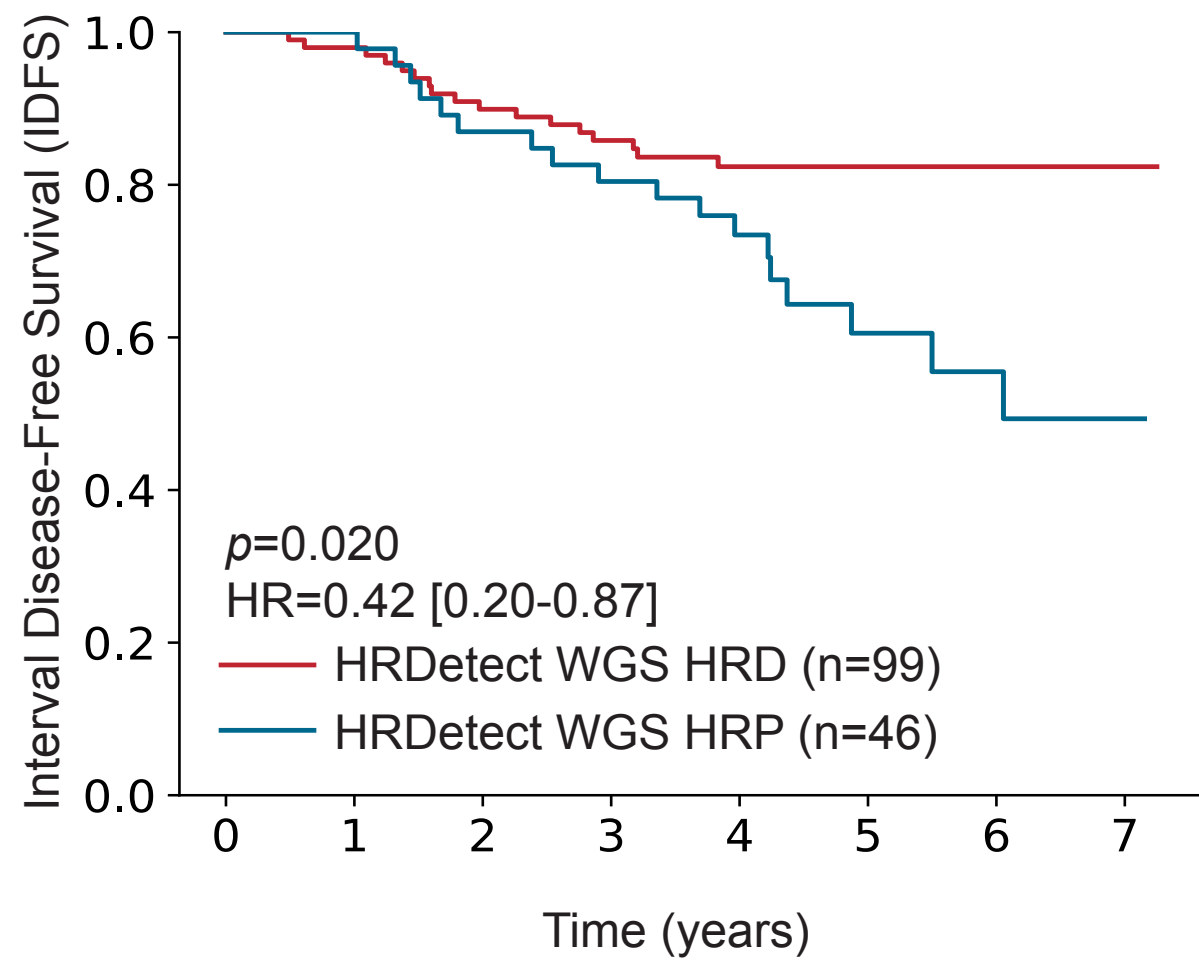
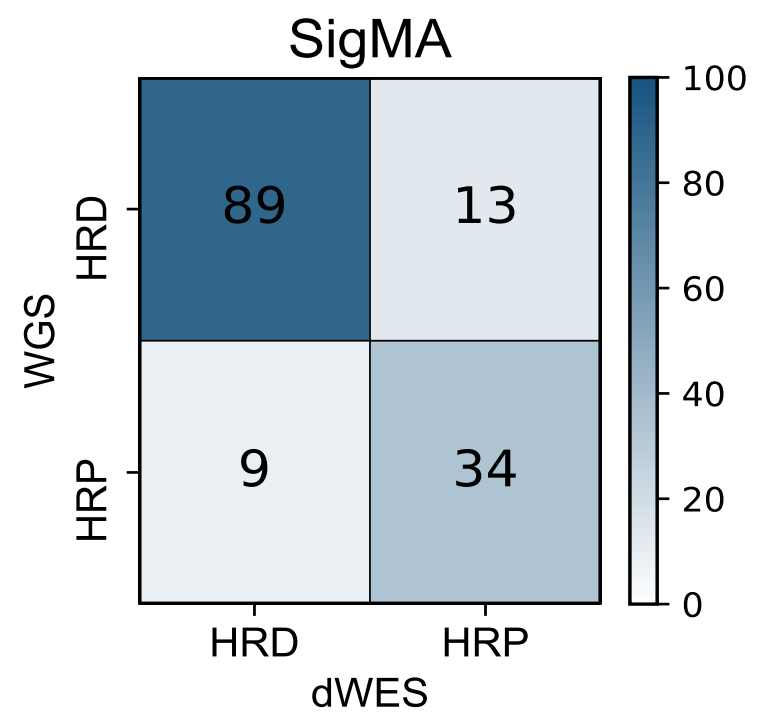
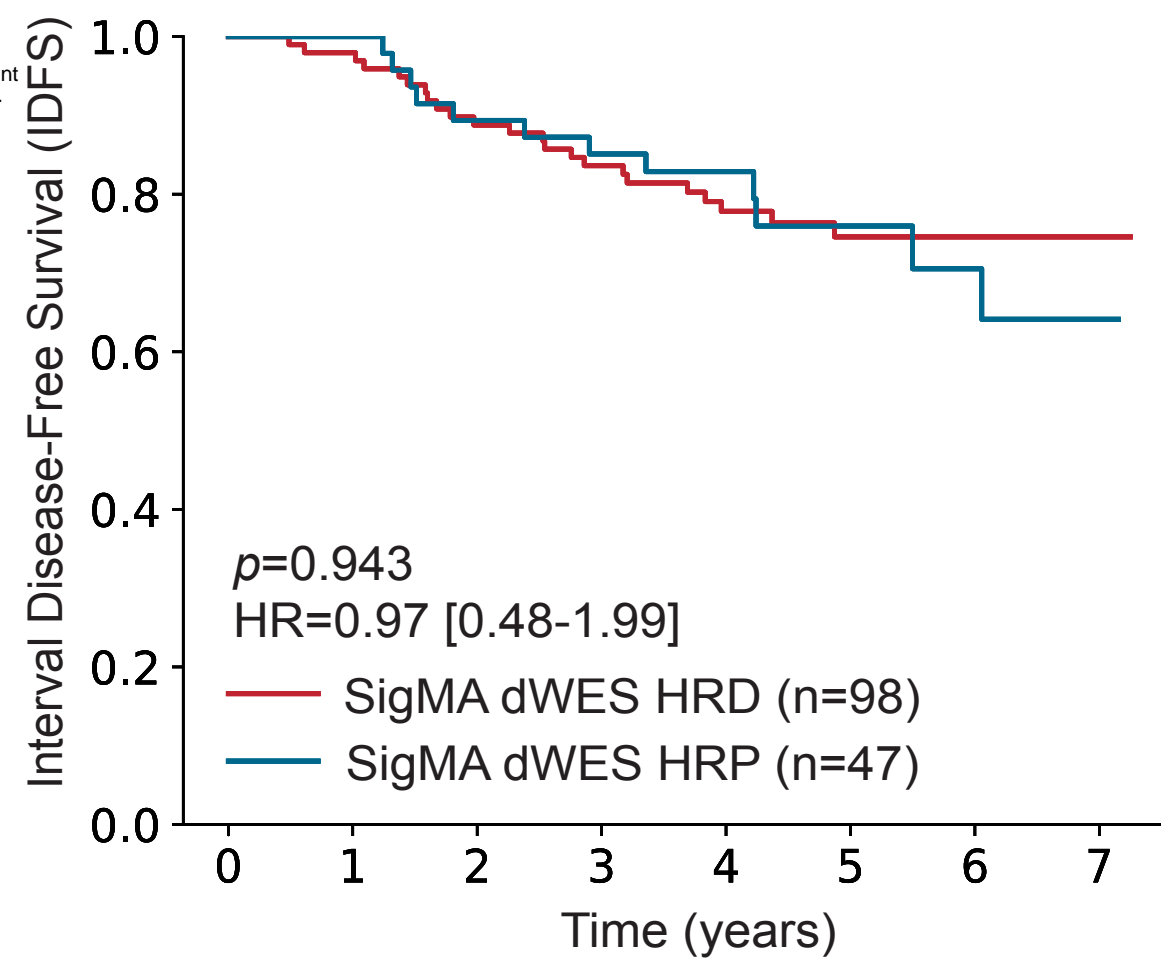
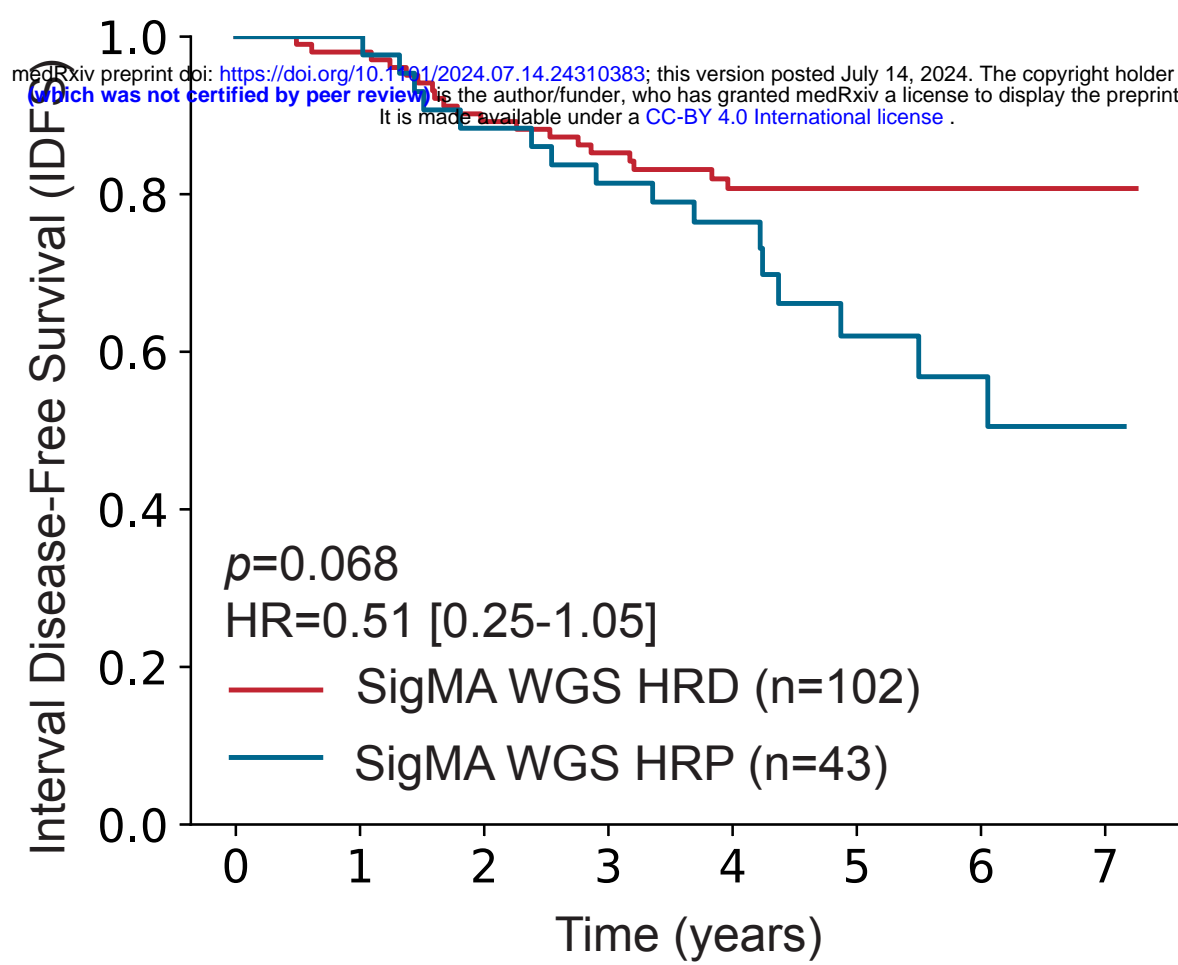
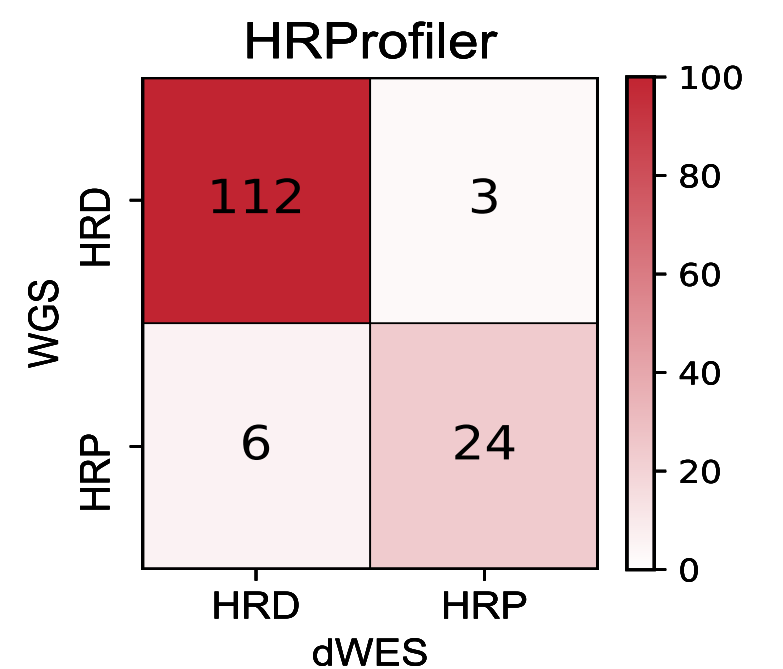
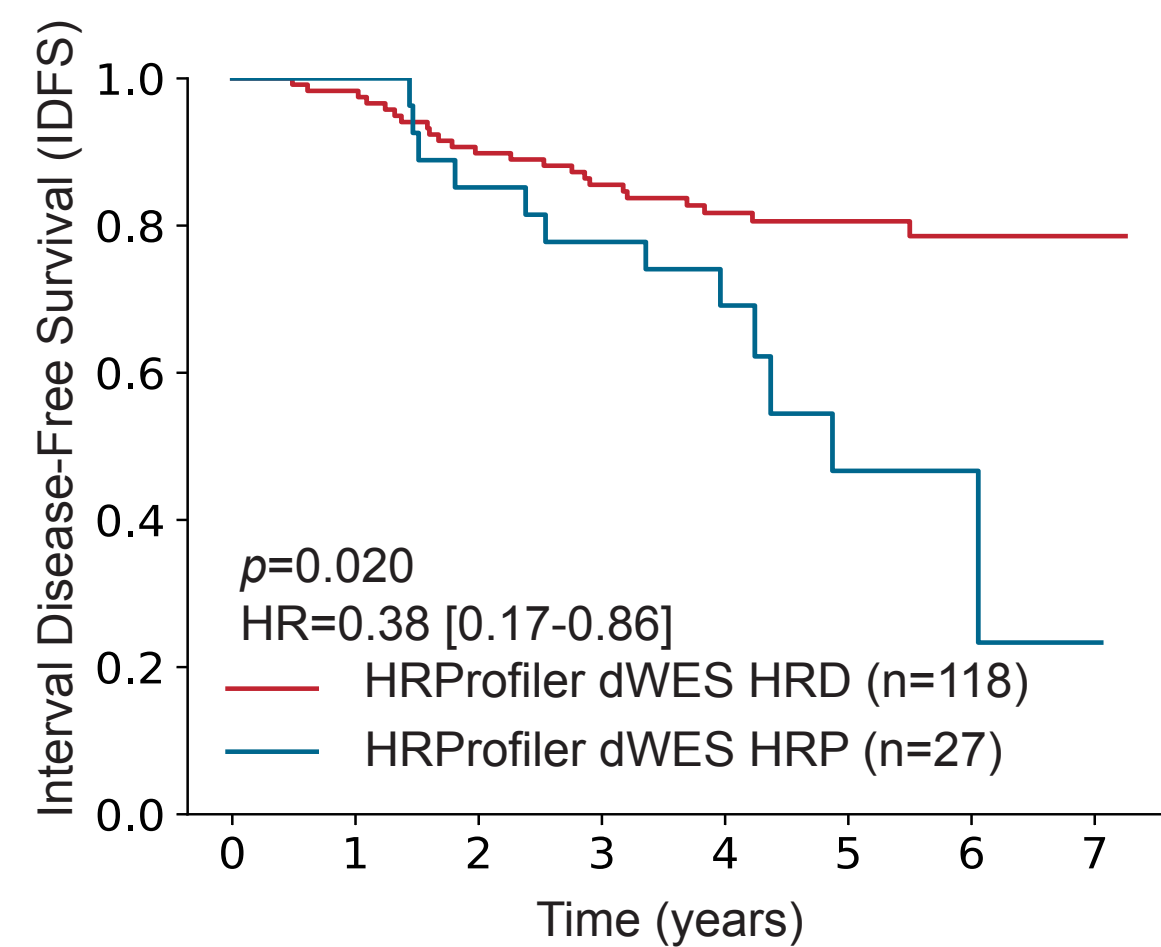
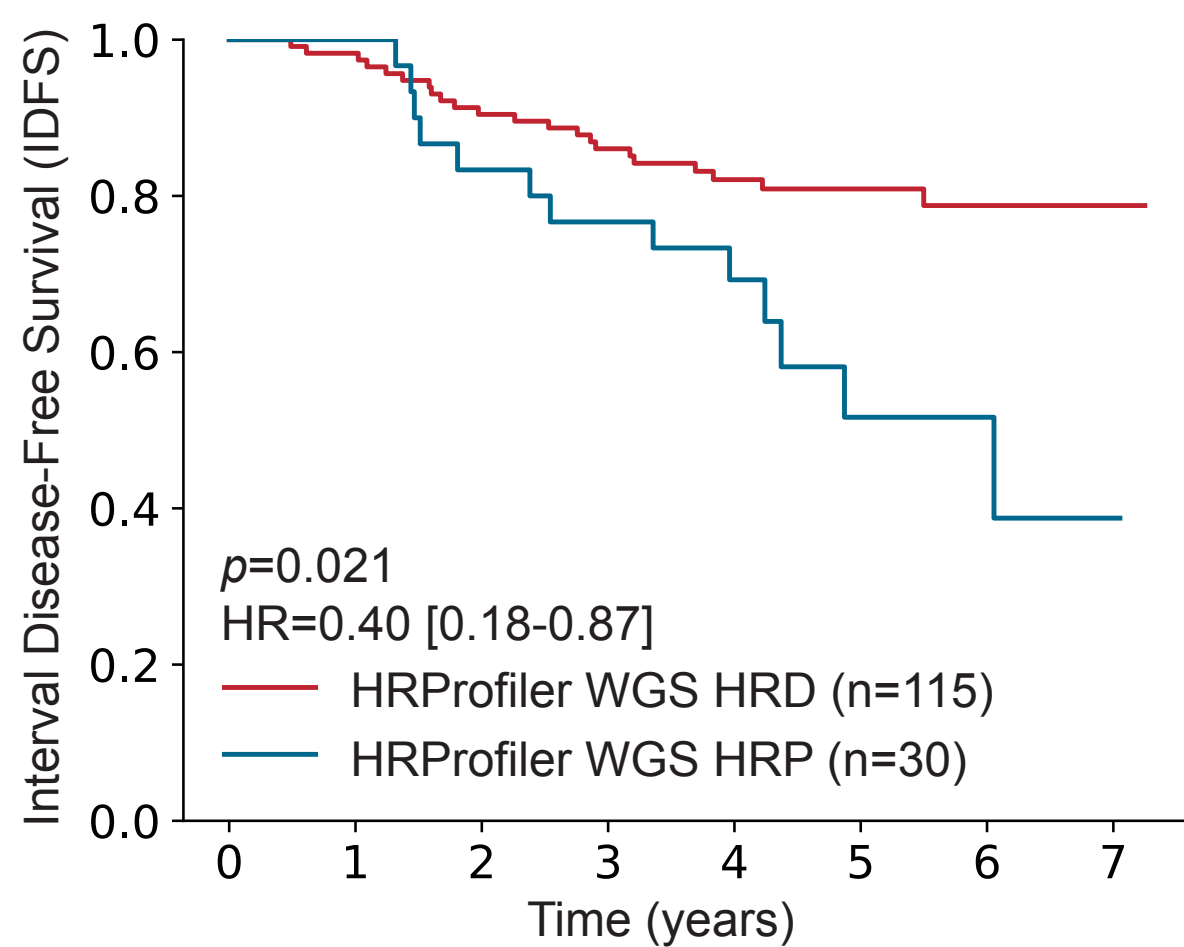
# Figure 1

It is made available under a [CC-BY 4.0 International license](https://creativecommons.org/licenses/by/4.0/).





**Figure 2****a.** 237 WGS Triple Negative Breast Cancers**b.** 71 WGS TCGA Breast Cancers**c.** 71 WES TCGA Breast Cancers**d.** 109 WES MSK-IMPACT Breast Cancers

**Figure 3****a.****HRDetect****b.****SigMA****c.****HRProfiler**

**Figure 4**

# Ovarian cancers with PARPi treatment

