Pan-cancer association of DNA repair deficiencies with whole-genome mutational patterns

Simon G. Sørensen[1,2], Amruta Shrikhande[3], Gustav A. Poulsgaard[1,2], Mikkel H. Christensen[1,2], Johanna Bertl[1,2], Eva R. Hoffmann[3]* and Jakob S. Pedersen[1,2,4]*

[1]Department of Molecular Medicine (MOMA), Aarhus University Hospital, Palle Juul-Jensens Boulevard 99, 8200, Aarhus N, Denmark.

[2]Department of Clinical Medicine, Aarhus University, Palle Juul-Jensens Boulevard 82, 8200, Aarhus N, Denmark.

[3]DNRF Center for Chromosome Stability, Department of Cellular and Molecular Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

[4]Bioinformatics Research Center (BiRC), Aarhus University, 8000, Aarhus C, Denmark.

*corresponding authors: jakob.skou@clin.au.dk, eva@sund.ku.dk

Abstract

Background

Cancer genomes are shaped by DNA damage and repair. DNA Repair deficiencies in cancers may result in characteristic mutational patterns, as exemplified by deficiency of *BRCA1/2* and efficacy prediction for PARP-inhibitors. We systematically evaluated the ability to identify and predict mono- or biallelic deficiencies of 736 DDR genes from associations with genome-wide mutational patterns, including structural variants, across 6,065 whole-genome sequenced cancers.

Results

We assembled 535 cancer-specific patient cohorts with shared deficiency of a DNA damage response gene. We trained and evaluated a predictive model for each cohort, and shortlisted 24 gene deficiencies that could be predicted with high accuracy from mutational patterns. These included expected gene deficiency models for *BRCA1/2*, *MSH3/6*, *TP53*, and *CDK12*. *CDK12* is associated with tandem-duplications, and we demonstrate that this association can predict gene deficiency with high accuracy (AUROC=0.97) in prostate cancers.

Our novel associations include deficiencies of *ATRX, IDH1, HERC2, CDKN2A, PTEN,* and *SMARCA4*. We found that loss of *ATRX* or *IDH1*, which are often co-mutated, were associated with the same mutational phenotype in central nervous system cancers; *HERC2* loss was associated with an increased number of short deletions in skin cancers and significantly co-occurring with *TP53* loss; and *PTEN* loss was associated with decreased levels of structural rearrangements in cancers of the uterus and the central nervous system.

Conclusion

Our systematic, generic approach yielded a catalogue of predictive models, which may provide targets for research and development of treatment, and potentially help guide therapy.

Keywords: DNA damage response, repair deficiency, predictive modeling, personalised medicine, mutational signatures.

Background

The DNA damage response (DDR) is a central component to the genetic integrity of cells, and deficiencies in DNA repair pathways may affect mutational patterns genome-wide(1–3). Furthermore, some DDR deficiencies are known to modulate the response to therapies, e.g., *BRCA1/2* deficiency renders cancers susceptible to treatment with PARP inhibitors(4), mismatch repair-deficient cancers are sensitive to checkpoint inhibitors(5), and *CDK12*-mutated cancers have a suggested sensitivity to CHK1 inhibitors(6). Because of this, efforts have been made to annotate inactivating mutations in DDR genes(7). However, the approach is limited by the lack of functional impact annotation of most variants, which are generally denoted as "variants of unknown significance" (VUS). Moreover, loss of gene activity could also occur by other means, such as transcriptional silencing.

A different approach is to investigate whether DDR deficiencies can be identified through associated genome-wide mutational scars. This approach has been pioneered for homologous recombination deficiency mediated by *BRCA1/2* deficiencies (-d), which may be successfully predicted by measuring the mutational accumulation of, primarily, small deletions flanked by microhomologous sequences(2,8).

Other mutational patterns are associated with specific DDR deficiencies and are thus potentially amenable to predictive modelling: Firstly, damage to the tumour suppressor gene *TP53* gene is related to increased genomic instability(9,10). Secondly, mismatch repair-deficient cancers, such as cancers with loss of *MSH3* or *MSH6*, are associated with microsatellite instability(11,12). Finally, *CDK12* deficiency is associated with a genome-wide phenotype of large tandem duplications(13,14). The scope of this approach can now be evaluated

73 systematically across DDR genes by exploiting available whole cancer genomes from
74 thousands of patients(15,16).
75
76 To initiate a systematic analysis of DDR genes, mutations observed genome-wide may be
77 condensed into mutational summary statistics for predictive modelling, including statistics based
78 on single-base substitutions (SBSs), indels, and different types of structural variants (SVs). The
79 SBSs may be statistically assigned to and represented by a set of so-called SBS signatures that
80 are catalogued and enumerated within the COSMIC database(17). Some of these are
81 associated with specific DNA repair deficiencies as well as genotoxic exposures, such as
82 ultraviolet (UV) light and smoking. Each mutational signature captures the relative frequency of
83 the different mutation types and their flanking nucleotides(18).
84
85 Here we performed a systematic screen for DDR gene deficiencies that can be predicted
86 through their association with genome-wide mutational patterns. We developed a generic
87 approach to train predictive statistical models that identify associations with individual mutational
88 summary statistics, including SBS signatures, indels, and large SVs. We applied it to 736 DDR
89 gene deficiencies, considering both monoallelic and biallelic loss of function (LOF), identified
90 across 32 cancer types, in a combined set of whole cancer genomes from 6,065
91 patients(15,16). The underlying aim was to identify novel associations with potential biological
92 relevance and to identify DDR deficiencies that can be predicted with sufficiently high certainty
93 to have a potential for clinical application.
94
95 Our analysis revealed 24 DDR genes where deficiencies are associated with mutational
96 summary statistics in individual cancer types across 48 predictive models. These results
97 recapitulated the expected associations between mutational patterns and deficiencies of

98 *BRCA1/2, TP53, MSH3/6,* and *CDK12*. We supplemented this knowledge by providing a
99 predictive model of *CDK12* deficiency that achieved high accuracy (AUROC=0.97) in prostate
100 cancer. Furthermore, we present unexpected predictive models of several DDR deficiencies;
101 *ATRX* and *IDH1* deficiency in cancers of the central nervous systems, *HERC2* and *CDKN2A*
102 deficiency in skin, *PTEN* deficiency in cancers of the central nervous system and uterus, and
103 *SMARCA4* deficiency in cancers of unknown primary.

104 Results

105 **DDR gene deficiencies across 6,065 whole cancer genomes**
106 We compiled and analysed 2,568 whole genome sequences (WGS) from The Pan-Cancer
107 Analysis of Whole Genomes (PCAWG)(15) and 3,497 WGS from the Hartwig Medical
108 Foundation (HMF)(16). In total we investigated 6,065 whole cancer genomes of 32 cancer types
109 (**Fig. 1a; Additional file 1, Supp. Table 1**).
110
111 For each genome, we evaluated 736 known DDR genes for both germline and somatic loss of
112 function (LOF) events(19–21). We annotated both mono- and biallelic LOF events, where each
113 event could be either a single nucleotide variant, an indel, or a loss of copy number (**Fig. 1b,c;**
114 **Additional file 1, Supp. Table 2**). Pathogenicity of SNVs and indels was evaluated using a
115 combination of CADD scores (>25; 0.3% most pathogenic) (22) and ClinVar annotation, when
116 available (**Methods**).
117
118 We inferred a total of 8,408 biallelic DDR gene deficiencies, primarily through a combination of
119 somatic or germline variants (SBSs and indels) with pathogenic potential (n=1,702), or loss-of-
120 heterozygosity (LOH) events combined with a single pathogenic germline (n=3,562) or somatic

121 (n=3,078) variant (SBS or indel; **Fig. 1b**). On average we observed a single, biallelic DDR gene
122 loss per patient, with some patients showing extreme rates of somatic pathogenic mutations
123 (**Fig. 1c**). As expected, *TP53* deficiency (*TP53*-d) was the most frequent LOF event (81 biallelic
124 and 1,746 monoallelic events; 29% of patients affected; **Additional file 1, Supp. Table 2; Fig.**
125 **1d**), while 70% of DDR genes had biallelic deficiency in less than ten samples across all cancer
126 types (511/736; **Fig. 1d**). Among monoallelic events, we identified 15,063 pathogenic germline
127 (59%) and 10,336 somatic (41%) events.
128
129 **Whole genome mutational summary statistics**
130 We collected mutational summary statistics for each cancer genome, which were used as
131 features for the downstream predictive models (**Fig. 1e,f; Additional file 2, Supp. Table 3-4**).
132 For SBSs, we evaluated exposure towards predefined sets of cohort-specific SBS
133 signatures(18,23). Short indels and SVs were simply categorised and counted: Deletions were
134 sub-categorised based on surrounding sequence repetitiveness and presence of
135 microhomology. SVs were sub-categorised by type (tandem duplications (tds), inversions (inv),
136 deletions (del), and translocations (trans)), five size ranges (not relevant for translocations), and
137 cluster presence (**Methods**). Several SBS signatures as well as some types of indels have
138 suggested etiologies (collected in **Additional file 1, Supp. Table 5)**.
139
140 **Statistical modeling of DDR gene deficiencies**
141 For the downstream statistical analysis we restricted our focus to DDR genes in cancer types
142 with more than five biallelic LOF events in either PCAWG or HMF (n=194 LOF-sets) or more
143 than ten monoallelic LOF events (n=341 LOF-sets) (**Additional file 1, Supp. Table 6**). Using
144 *BRCA2*-d in the set of HMF breast cancer samples (n=645) as an example, we observed
145 biallelic LOF events in 17 (2.6%; 14 germline, 3 somatic) patients and monoallelic LOF events in

146 7 (1.1%; 4 germline, 3 somatic) samples (**Additional file 1, Supp. Table 8**). We further
147 observed VUS events in 53 patients (8.2%; 42 germline, 11 somatic), which were excluded from
148 the analysis. The remaining *BRCA2* wild-type (WT) samples (n=568; 88.1%) were used as a
149 background set for training the predictive models (**Fig. 2a)**.
150
151 For each LOF-set we trained a LASSO logistic regression model and evaluated the ability to
152 discriminate between deficient and WT samples (**Methods**). For *BRCA2*-d we observed a
153 strong association with the number of deletions at sites of microhomology (**Fig. 2b**), with a
154 median of 608 deletions per patient in *BRCA2*-d breast cancers versus 81 in *BRCA2* WT breast
155 cancers, in agreement with prior findings(2,8,24). The LASSO regression also included non-
156 clustered inv 10-100Kb and clustered tds 1−10Kb, although both show high variance among
157 patients for both deficient and WT (**Fig. 2b**) and have considerably smaller coefficients,
158 ultimately contributing little influence on overall predictive performance (**Fig. 2c**).
159
160 For each LOF-set we evaluated the predictive performance using the area-under-the-receiver-
161 operating-curve (AUROC) score as well as the precision-recall area-under-the-curve (PR-AUC)
162 score. The PR-AUC score is a more robust measure for unbalanced data sets(25); however, the
163 expected value for non-informative (unskilled) models equals the fraction of true positives and
164 thus varies between LOF-sets. Therefore, we used the PR-AUC enrichment over the
165 expectation (PR-AUC-E) as our selection criteria for predictive models. For the downstream
166 analysis, we shortlisted and only considered models found to have a PR-AUC-E that was
167 substantial (>0.2; more than two standard deviations above the mean across all 535 LOF-sets)

168 and significant (Benjamin-Hochberg False discovery rate (FDR)<0.05; Monte Carlo simulations)
169 (**Methods**; **Additional file 1, Supp. Table 7**).
170
171 In the example of *BRCA2*-d in breast cancers of the HMF data set, our model achieved an
172 AUROC of 0.93 and a PR-AUC-E of 0.29 (**Fig. 2d,e; Additional file 1, Supp. Table 7**). Out of
173 30,000 permuted LOF-sets, none had a similar or higher PR-AUC score and we considered the
174 model significant with a p-value < $3x10^{-5}$ (FDR adjusted q-value < $6x10^{-4}$) (**Fig. 2f,g**). Notably,
175 the *BRCA2*-d model did not include non-clustered deletions <100Kb, SBS signature 3, and SBS
176 signature 8, all features which have been associated with BRCAness(8). However, SBS
177 signature 3 and non-clustered deletions 1-10Kb in size are included in the model when the
178 deletions at sites of microhomology are omitted from the input data set, suggesting that they are
179 excluded during feature selection due to their high positive correlation with the number of
180 deletions at sites of microhomology among HMF breast cancers (Pearson corr. > 0.7; **Fig. 2h**;
181 **Additional file 1, Supp. Table 9**).
182
183 Our selection criteria resulted in 48 shortlisted predictive models across 24 DDR genes (**Fig. 3a;**
184 **Additional file 1, Supp. Table 7)**. As exemplified for *BRCA2*, each model is specified by a set
185 of predictive features representing mutational patterns associated with DDR gene LOF**.** We
186 divided the models into four groups based on etiology and origin: models of *BRCA1/2*-d (eight
187 models of *BRCA2*-d and a single model of *BRCA1*-d; **Fig. 3b**); models of monoallelic *TP53*-d
188 (11 models; **Fig. 3c**); models of various monoallelic gene deficiencies derived from colorectal
189 cancer patients (eight models; **Fig. 3d**); and models including other DDR genes and cancer
190 types, including previously undescribed associations (20 models; **Fig. 3e**).
191

192 **Recapitulation and predictive modelling of expected associations**

193 ***BRCA1/2* deficiency**

194 Five models predicted biallelic loss of *BRCA2* in cancers of the ovary, prostate, pancreas, and
195 breast. In addition, three models predicted *BRCA2* monoallelic loss in cancers of the pancreas,
196 breast, and prostate. Finally, we derived a single model of biallelic *BRCA1* loss in ovarian
197 cancer (**Fig. 3c; Fig. 4a**). All models significantly outperformed their Monte-Carlo simulations
198 ($q<0.05$; Benjamini Hochberg FDR control) and had PR-AUC-E above 0.2 (**Fig. 4b,c**).

199

200 All *BRCA2*-d models were predominantly predicted by deletions at sites of microhomology,
201 consistent with the role of *BRCA2* in homologous recombination and suppression of
202 microhomology-mediated end joining(26). Both clustered and non-clustered tds in the range of
203 1-100Kb were included as features for various models, though with much smaller predictive
204 power. The biallelic breast cancer model based on PCAWG further included SBS signature 3 as
205 a predictive feature (**Fig. 4d**). Contrasting to the models of *BRCA2*-d, *BRCA1*-d in ovarian
206 cancer was exclusively associated with clustered and non-clustered tds (1-10Kb; **Fig. 4d**). This
207 aligns with the findings of Davies *et al*.(8)**,** which also found that *BRCA1*-d was more closely
208 associated with a tandem-duplicator phenotype.

209

210 ***TP53* deficiencies**

211 We detected 11 predictive models (four based on PCAWG and seven on HMF) of monoallelic
212 *TP53*-d across cancers of the breast, skin, ovary, uterus, neuro-endocrine tissues (NET), biliary
213 gland, head and neck, pancreas, and the central nervous system (CNS) (**Fig. 4e**). These
214 predictive models performed with PR-AUC-E values ranging from 0.21 to 0.48 in breast and
215 biliary gland cancers, respectively. Similarly, AUROC values ranged from 0.48 to 0.88, again in
216 breast and biliary gland cancers (**Fig. 4f,g**). In line with existing literature(27), *TP53-d* is

217 associated with a significantly increased number of structural variants across the genome
218 (Wilcoxon rank-sum test; **Fig. 4h,i**), except in skin cancers.
219
220 **Colorectal cancer models**
221 In the HMF colorectal cancers, we discovered eight predictive gene deficiency models (*MSH3,*
222 *SMC2, SMC6, BMPR2, CLASP2, SRCAP*, *UBR5,* and *UVRAG)* of monoallelic LOF with PR-
223 AUC-E ranging from 0.21 (*CLASP2*) to 0.62 (*MSH3*) (AUROC ranging from 0.68 for *SRCAP*-d
224 to 0.94 for *MSH3*-d; **Fig. 4j,k,l**).
225
226 The high number of models of monoallelic deficiencies may arise from spurious LOF events in
227 DDR genes in a subset of colorectal cancers that are hypermutated. Some colorectal cancers
228 are signified by mismatch repair deficiencies, such as LOF of *MSH3* or *MSH6*, creating a high
229 number of deletions in repetitive DNA(11,12). Indeed, we found that this pattern was most
230 profound among the *MSH3* mutated cancers (**Fig. 4m,n**). Furthermore, we found co-mutation
231 with *MSH3* across all LOF-sets ranging from 20% (*SRCAP*) to 33% (*BMPR2*) of the patients
232 with mutations (**Fig. 4p**), suggesting that the models (except for the model of *MSH3*-d) might be
233 the consequence of the hypermutator phenotype. In other words, the causality is reversed in
234 these cases, and the mutational process driven by *MSH3*-d caused the majority of their LOF
235 events. This notion is supported by investigating the features of the models. All eight models are
236 characterized by a single, primary predictive feature: Insertions (*SMC2, SMC6, BMPR2, UBR5,*
237 *and UVRAG*), deletions in repetitive DNA (*MSH3, SRCAP*), or deletions not flanked by repetitive
238 or microhomologous DNA (*CLASP2*) (**Fig. 4o**). Each of these features has a high positive
239 correlation (Pearson corr. > 0.93) with the number of deletions in repetitive DNA. This
240 correlation suggests that all eight models relate to a genome-instability phenotype, which may

241 be driven by the *MSH3* co-mutated patients or, potentially, a concurrent deficiency of other
242 genes within the mismatch repair system (**Additional file 1, Supp. Table 9**).
243
244 **Biallelic LOF of *MSH6* associated with increased number of deletions in repetitive DNA in**
245 **prostate cancer**
246 *MSH6*, a gene implicated in mismatch repair and microsatellite stability(12)**,** was mutated in
247 both alleles in 7 out of 342 HMF prostate cancer patients. Deficiencies could be predicted with
248 high accuracy (PR-AUC-E=0.25; AUROC=0.98) by an enrichment of deletions in repetitive DNA
249 (**Fig. 5a,b; Additional file 1, Supp. Table 7**). This is consistent with existing findings of
250 mismatch repair deficiency and its presence in metastatic prostate cancers(28).
251
252 **High predictive power for *CDK12* deficiency**
253 *CDK12* is a kinase that regulates transcriptional and post-transcriptional processes of the
254 DDR(29–31)**.** We found that *CDK12*-d prostate cancers had an increased number of mid- and
255 large-sized tandem duplications (100Kb - 10Mb), compared to *CDK12*-WT (**Fig. 5c,d**). Several
256 studies have observed similar tandem duplication phenotypes in ovarian cancers(13,14,32) and
257 castration resistant prostate cancers(33,34). In agreement, nine of the ten patients in our data
258 set were treated with drugs associated with castration resistance (4 Enzalutamide, 3
259 Abiraterone, 1 Caba.,1 Pembro.)(35).
260
261 Whereas the tandem duplication phenotype has been previously associated with loss of *CDK12*,
262 in this study we present the first predictive algorithm utilising and quantifying the high predictive
263 value of these patterns (PR-AUC-E=0.73 and AUROC=0.97). Indeed, the loss of CDK12 has
264 been demonstrated to sensitise cancer cells to CHK1-(6) and PARP-inhibitors(36,37).
265

266 We went on to test the *CDK12*-d model across other cancer types (**Fig. 5c**). As expected, we
267 observed predictive power in cancers of the ovary and breast, though at a lower level (PR-AUC-
268 E=0.19 and AUROC=0.72). No predictive power was observed for the remaining cancer types.
269
270 **Novel predictive gene deficiency models**
271 The shortlisted predictive LOF models further include biallelic LOF of *ATRX, PTEN, HERC2,*
272 *MEN1, SMARCA4, BAP1,* and *RB1,* as well as monoallelic LOF of *ATRX, IDH1, PTEN,*
273 *CDKN2A, ARID1A, TP53BP1, HERC2,* and *RB1* (**Fig. 3e; Additional file 1, Supp. Table 7)**.
274 The number of mutated patients in each LOF-set varied from eight (biallelic *SMARCA*-d in
275 cancers of unknown primary) to 22 (monoallelic *HERC2*-d in skin). We here present the models
276 of *ATRX, IDH1, SMARCA4, CDKN2A, PTEN, and HERC2*; remaining models may be found in
277 the supplementary material (**Additional file 1, 3, 4, Supp. Table 7, Supp. Fig. 1**).
278
279 **Predictive models of *ATRX*-d and *IDH1*-d in CNS cancers**
280 We found that *ATRX*-d (monoallelic: PR-AUC-E=0.21, AUROC=0.71; biallelic: PR-AUC-E=0.23,
281 AUROC=0.76; **Fig. 6a**) and *IDH1*-d (monoallelic: PR-AUC-E=0.24, AUROC=0.82; **Fig. 6b**) in
282 CNS cancers were both predicted by a decreased number of SBS signature 8 mutations. In
283 addition, *ATRX*-d was further predicted by non-clustered inv. 10-100Kb, although with a small
284 coefficient and hence contributing limited discriminatory power. We discovered that 7 of11 *IDH1*
285 mutated patients were also *ATRX* mutated (28-fold enrichment compared to *IDH1* WT skin
286 cancers; $p=1.2x10^{-8}$, Fisher's exact test).
287
288 Co-mutation between *IDH1* and *ATRX* is well-described in gliomas(38). Interestingly, LOF of
289 either gene is associated with lack of SBS signature 8, a signature associated with

290 BRCAness(8,39) and late-replication errors(40). This suggests that CNS cancers with
291 *IDH1/ATRX* deficiency are not subject to the same mutational processes as other CNS cancers.
292
293 **Predictive model for *SMARCA4*-d in cancer of unknown primary**
294 We discovered eight patients (HMF) out of 77 with cancers of unknown primary with *SMARCA4*-
295 d (biallelic) that could be predicted with relatively high accuracy (PR-AUC-E=0.44;
296 AUROC=0.85; **Fig. 6c**). These cancers showed an enrichment of SBS signature 27 (a signature
297 first detected in myeloid cancers(41)), which has been suggested to be a sequencing artefact
298 though it also displays strong strand bias(17) (**Additional file 1, Supp. Table 5**). Among
299 cancers of unknown primary, SBS signature 27 correlates strongly with SBS signature 4
300 (Pearson corr.=0.96), despite the signatures different composition (cosine similarity of
301 signatures = 0.17) (**Additional file 1, Supp. Table 9; Fig. 6h,i**). This suggests that SBS
302 signature 4 may also be associated with *SMARCA4*-d and we indeed found that its predictive
303 performance (PR-AUC-E of 0.43; AUROC=0.83) was almost equivalent to SBS signature 27.
304 Signature 4 is associated with smoking across several cancer types(42,43); interestingly,
305 *SMARCA4*-d is seen in aggressive thoracic sarcomas(44) and strongly enriched among patients
306 with a history of smoking(45). Its gene product, BRG1, has been suggested as a lung cancer
307 transcriptional regulator of genes that induce tumour proliferation(46) and metastasis(47). We
308 evaluated the ability of SBS signature 4 to predict *SMARCA4*-d in other cancer types and found
309 a significant predictive association in lung cancer, though much lower than for cancers of
310 unknown primary (**Fig. 6j**). We also found a significant ability to predict *SMARCA4*-d by the
311 number of SBS signature 27 mutations in cancers of the NET (Wilcoxon test, one-tailed
312 p=0.002) and head and neck ($p=4.8x10^{-7}$), but we could not evaluate this in lung cancers, as the
313 signature is not among its set of cohort-specific signatures.
314

315 Twelve cancers had a high posterior probability of *SMARCA4*-d (**Fig. 6j**) despite being
316 *SMARCA4* WT and lacking pathogenic events. No other single DDR gene was mutated among
317 all twelve patients, with *TP53* having the most LOF events (6 of 12 cancers). Given that cancers
318 of unknown primary have disparate origins, this raises the possibility that the patients with high
319 posterior probability of *SMARCA4*-d may have metastasized from a cancer type or subtype with
320 both high levels of SBS signature 4 and 27, as well as high incidence of *SMARCA4*-d. Further
321 studies are thus needed to clarify if the observed associations can be explained through such
322 an ascertainment bias rather than causatively by *SMARCA4*-d.
323
324 **Predictive model for monoallelic *CDKN2A*-d in skin cancer**
325 Both germline and somatic variants in *CDKN2A* are known to predispose for melanoma(48). In
326 the PCAWG skin cancer cohort, we found that a monoallelic predictive model of *CDKN2A*-d
327 achieved relatively high accuracy (PR-AUC-E=0.28; AUROC=0.82). Its predictive features are
328 enrichment of deletions flanked by microhomology, non-clustered inv. (100Kb-1Mb), and SBS
329 signature 7, in order of predictive importance (**Fig. 6d**). Apart from the inversions, these features
330 are also significantly associated with biallelic *CDKN2A*-d in the HMF skin cancer cohort
331 (**Additional file 4, Supp. Fig. 2a,b**) and are included in the corresponding predictive model,
332 though it had lower predictive performance (PR-AUC-E=0.008; AUROC=0.542) and was not
333 shortlisted (**Additional file 1, Supp. Table 6**). In both HMF and PCAWG, most of the observed
334 LOF events are somatic, with the majority being biallelic (16 of 18 in PCAWG; 32 of 33 in HMF;
335 **Additional file 1, Supp. Table 8**). The presence of deletions flanked by microhomology
336 suggests a possible reduction in error-free double-stranded break repair in combination with an
337 increased accumulation of SBS signature 7.
338

339 **Predictive model for *HERC2-d* in metastatic skin cancer**

340 *HERC2*-d has been associated with susceptibility to developing melanoma(49). We found
341 *HERC2*-d predictable in HMF skin cancer patients (biallelic: PR-AUC-E=0.24, AUROC=0.73;
342 monoallelic: PR-AUC-E=0.23, AUROC=0.66; **Fig. 6e)**, primarily based on enrichment of
343 deletions unflanked by microhomology or repetitiveness.

344

345 *HERC2* encodes a protein ligase that modulates the activity of P53(50). We observed that
346 seven of the nine patients with biallelic *HERC2*-d also had a monoallelic pathogenic event in
347 *TP53* (8-fold enrichment; $p=4.6x10^{-6}$, Fisher's exact test). Patients that are co-mutated in *TP53*
348 (monoallelic) and *HERC2* (mono- or biallelic) showed a significantly higher number of deletions
349 compared to patients with LOF in either gene alone (Wilcoxon test, One-tailed $p<0.031$) and
350 cancers that are wild-type for both genes ($p<7.2x10^{-7}$) (**Additional file 4, Supp. Fig. 2c**).

351

352 ***PTEN* deficiency associates with fewer structural variants in CNS and uterine cancers**

353 *PTEN* is a tumour suppressor gene found in various cancer types(51,52) and its deficiency was
354 found to be associated with mutational patterns in uterine and CNS cancers (**Additional file 1,**
355 **Supp. Table 7**). In CNS, we acquired two identical models from the HMF data set, as we
356 observed no monoallelic events without a second hit (mono- and biallelic PR-AUC-E=0.37,
357 AUROC=0.74; **Fig. 6f**). In uterine cancer, we acquired significant predictive models from both
358 HMF and PCAWG (HMF, biallelic: PR-AUC-E=0.37, AUROC=0.74; HMF, monoallelic: PR-AUC-
359 E=0.36, AUROC=0.79; PCAWG, monoallelic: PR-AUC-E=0.22, AUROC=0.74; **Fig. 6g**). The
360 model of biallelic *PTEN*-d in uterine cancers is primarily based on depletion of non-clustered inv.
361 (10-100Kb), whereas both the HMF and PCAWG models of monoallelic deficiency are primarily
362 based on depletion of non-clustered tds. (10-100Kb) (**Fig. 6g**). In contrast, our models of

363 *PTEN*-d in CNS cancers from HMF were based on depletion of both non-clustered deletions (1-
364 10Kb) and non-clustered translocations (**Fig. 6f**).

365 Discussion

366 The cancer-specific, incomplete repair of endogenous and exogenous DNA lesions leave
367 specific genome-wide mutational patterns. Their detection provides potentially powerful
368 information on the fidelity of individual DNA repair pathways and response to chemo- and
369 immunotherapies. Taking advantage of a large pan-cancer data set, our analysis shows that
370 mutational patterns associate with DNA repair defects across a wide range of cancers and
371 repair mechanisms. In this study, we have contributed concrete initial predictive algorithms for
372 mutational patterns of several DDR gene deficiencies, with potential use for clinical intervention.
373
374 The clinical scope of our approach is exemplified by recent regulatory approval for PARPi
375 administration, which was supported by statistical predictions of homologous recombination
376 deficiency (HRD) based on mutational patterns(53) We have similarly predicted *CDK12*
377 deficiency (*CDK12*-d) which has previously been associated with a tandem duplication
378 phenotype– with power similar to that of HRD detection. This suggests a clinical benefit of
379 clinical application of predictive algorithms to test for *CDK12*-d, specifically when considering
380 the application of CHK1-inhibitors in prostate, ovary, and breast cancer treatment(6).
381
382 We also find high predictive power for *MSH6*-d in prostate cancers by counting deletions in
383 repetitive DNA, also known as the microsatellite instability phenotype (MSI)(54). It is common
384 practice to search for signs of MSI in colorectal cancers, endometrial cancers, and aggressive
385 prostate cancers. This prediction is most commonly made using a panel of repetitive DNA

386 regions, such as the Bethesda panel(55). Here we demonstrated high predictive power in
387 prostate cancers based on mutation summary statistics, which may be routinely extracted from
388 whole genome sequencing. The correct identification of patients with high MSI supports the
389 administration of immune checkpoint blockade treatment, as the mutational phenotype leads to
390 an increased expression of neo-peptides and thus an increased sensitivity towards the immune
391 response(56,57).
392
393 Recent studies have suggested the use of CDK4/6 inhibition in *SMARCA4* deficient
394 patients(58). We discovered that a large subset of cancers of unknown primary (>10%) have
395 *SMARCA4* deficiency and that these cancers can be accurately predicted from their whole
396 genome accumulation of SBS signature 4 or SBS signature 27 mutations. This is a clinically
397 challenging cancer type due to the lack of a primary cancer to guide treatment and any such
398 prediction may serve in selecting treatment for these complex cancers; however, additional
399 studies are needed to establish causality.
400
401 Of note, the inclusion of monoallelic events allowed for larger sets of patients with DRR gene
402 LOF, thereby increasing the power of our study when causative associations with mutational
403 patterns exist. This is for instance the case for *TP53,* which is known to be functionally impacted
404 by monoallelic events(59). We shortlisted 11 models of TP53-d, which were all monoallelic. We
405 further found that inclusion of monoallelic LOF events matched the predictive power for *ATRX*
406 biallelic LOF in CNS cancers and *HERC2* biallelic LOF in skin cancers, suggesting that single
407 genetic hits are sufficient to affect repair and ensuing mutational patterns for these genes.
408
409 Incorrectly annotated LOF events in DDR genes may affect both the ability to discover
410 associations with mutational patterns and the performance of any associated predictive models.

411 On the other hand, highly stringent LOF criteria may result in true LOF events being overlooked,
412 and create small deficiency sets with insufficient power to discover true associations. For the
413 annotation of germline LOF events, we restricted our focus to rare variants (population
414 frequency ≤0.5%), as more common variants are likely benign. Yet, given the high number of
415 variants evaluated overall, false positive germline LOF calls are expected. However, they are
416 unlikely to associate with specific mutational patterns and thus are unlikely to contribute
417 significant, shortlisted false positive predictive models of DDR deficiency. Likewise, in cancers
418 with hypermutator phenotypes, somatic LOF events may be caused by and associated with
419 specific mutational patterns. This leads to predictive models that in effect detect instances of
420 reverse causality, as discussed for the monoallelic models found in colorectal cancer. We
421 investigated whether reverse causality was a likely explanation for an association captured by a
422 predictive model by evaluating whether the annotated LOF events matched the predictive
423 mutational features (**Additional file 1, Supp. Table 9**). In general, biallelic LOF criteria is less
424 sensitive to wrongly annotated LOF events than monoallelic criteria, as double hits are rare
425 compared to single hits. For both mono- and biallelic predictive models, the data permutations
426 ensure that the observed association between the DDR gene LOF events and given mutational
427 patterns are surprising.
428
429 Our systematic approach provides a proof-of-concept that will become increasingly powerful as
430 the available data sets increase in number and size. For the current data sets, consistent
431 validation of detected associations was challenging due to small cohorts and differences in
432 cancer biology.
433

Conclusions

The current study yields a catalogue of predictive models that captures both known and novel associations between DDR gene deficiencies and mutational patterns. The included DDR genes may provide targets for research and development of treatment. With further optimization, predictive models such as these may guide the selection of therapy by adding certainty of disabled or compromised repair deficiency phenotypes found in cancers of individual patients.

Methods

Data

The analysis was conducted on 6,098 whole cancer genomes filtered down to a set of 6,065 whole cancer genomes (see below). The data came from two independent data sets: the Pan-Cancer analysis of Whole Genomes (PCAWG; #Samples = 2,583; ICGC study ID. EGAS00001001692)(15) and the Hartwig Medical Foundation (HMF; #Samples = 3,515; Acc. Nr. DR-044)(16). The data sets contain patients across 32 primary tumour locations and represent diverse patient groups in terms of age, gender, and disease history. PCAWG consists of tumours both with and without metastasis, whereas HMF exclusively consists of donors with tumours showing metastatic capability. In order to best relate the two data sets, all tumours are catalogued by the site of primary disease. In 77 HMF samples, the primary site was unknown and these are annotated as such (**Additional file 1, Supp. Table 1**).

454 **Filtering variants for the annotation of LOF in DDR genes**

455 We annotated variants across a set of 736 genes related to the DNA damage response
456 (**Additional file 1, Supp. Table 2**). The set of genes is combined from three sources,
457 Knijnenburg et al.(20); Pearl et al.(19) and Olivieri et al.(21).

458

459 We filtered the downloaded variant call files (VCF) of all samples for variants between the
460 Ensembl (GRCH37/hg19) start- and end position of each DDR gene (see coordinates in
461 **Additional file 1, Supp. Table 2**). We discarded variants classified as none-PASS in the VCF
462 files and variants of the PCAWG data set supported by only one of our four variant callers.
463 Furthermore, we discarded all variants which occurred in more than 200 samples across the two
464 data sets in order to avoid noise arising from single nucleotide polymorphisms (SNPs) called as
465 single nucleotide variants (SNVs), SNPs with high frequency in particular populations and
466 possibly technical artefacts. We also discarded somatic variants with variant allele fractions
467 (VAF) below 0.2 and variants where gnomAD (V2.1.1) showed a germline population frequency
468 above 0.5%(60).

469

470 **Filtering samples**

471 In the HMF data, we selected the earliest available sample whenever multiple samples existed
472 for the same patient. For the PCAWG data, we used the official whitelist, obtained at the ICGC
473 resources (https://dcc.icgc.org/pcawg), and for patients with more than one sample, we always
474 selected the earliest, whitelisted sample available. We discarded six PCAWG samples of
475 bone/soft tissue, as these were discarded in prior publications based on the PCAWG data
476 set(23) leaving 2,568 PCAWG samples and 3,497 HMF samples across 32 sites of the body. A
477 list of all Donor IDs, Sample IDs, and primary sites of disease may be found in **Additional file 1,**
478 **Supp. Table 1**.

479

480 **Annotating loss of heterozygosity and deep deletions**

481 We also used the copy number profiles of each sample to discover genes with biallelic or
482 monoallelic losses of parts of genes. Any overlap between a gene and a copy number loss was
483 indicated as a loss-of-function event, under the assumption that losing any part of the protein
484 coding DNA is detrimental to the complete protein product. We considered events in which the
485 minor allele copy number was below 0.2 to be a loss of heterozygosity, whereas we considered
486 events with a total tumor copy number below 0.3 (major and minor allele summarized) as deep
487 deletions. This cutoff is adapted from the work of Nguyen et al.(24).

488

489 **Annotating pathogenic variants and mutations**

490 We annotated all variants and mutations with CADD phred scores (V1.6) in order to separate
491 likely pathogenic variants from likely benign variants(22). Furthermore, we annotated all variants
492 with their status in the ClinVar database (when present). Combining ClinVar and CADD phred
493 scores ensured that we would be able to discover associations across all DDR genes, not only
494 genes with ClinVar annotations, which are expected to be incomplete for any given gene.

495

496 All variants with CADD phred scores of 25 or higher were considered pathogenic, whereas non-
497 synonymous mutations with CADD phred scores below 25 and above 10 were considered
498 variants of unknown significance (VUS). Variants with CADD phred scores below 10 were
499 considered benign. With the cutoff value of 25, we only consider variants estimated to be among
500 the 0.3% most pathogenic.

501

502 **Annotating biallelic and monoallelic hits**

503 We considered genes with a single pathogenic variant, either somatic or germline, to be

504 monoallelic hit. In cases where the monoallelic hit is accompanied by a loss-of-heterozygosity
505 (LOH), we considered the event a biallelic loss. Finally, genes that were completely depleted in
506 a sample were considered biallelic lost. We did not consider a single LOH event as a
507 monoallelic loss due to the high frequency of such events, with ~25 times higher rates of LOH
508 than deep deletions.
509
510 **Mutational summary statistics of single base substitutions, indels and structural variants**
511 We summarised the genome-wide set of somatic mutations by using Signature Tools Lib,
512 developed by Degaspari et al(23).
513
514 Firstly, we used Signature Tools Lib to count single nucleotide variants by base change and
515 context, and assigned these counts to a set of organ-specific signatures, as recommended by
516 Degaspari et al. To ensure the ability to test models across cancer types, we have converted
517 the organ-specific signature exposures to reference signature exposures using the conversion
518 matrix found in **Additional file 1, Supp. Tables 9, 10** in the paper by Degaspari et al(23). For
519 twelve signatures (N1-N12) this was not possible, and these signatures are considered to be
520 novel. Likewise two signatures associated with mismatch repair (MMR) deficiency, MMR1 and
521 MMR2, associated with several of the COSMIC MMR signatures. For these, we preserved the
522 signature label as given by signature tools lib. For cohorts with no organ-specific signatures
523 (cancers of unknown primary, neuroendocrine tissue, thymus, urinary tract, penis, testis, small
524 intestine, vulva, double primary, and the adrenal gland), we have assigned mutations directly to
525 the full set of 30 reference signatures (COSMIC signatures 1-30). Note that we excluded age-
526 related signature 1 from the modeling, as this signature confounded by acting as a proxy for the
527 age of the patient so that the model would be learning to differentiate old from young patients
528 rather than patients with activity of different mutational processes or repair deficiencies.

529

530 Secondly, Signature Tools Lib was used to count indels: insertions were simply counted,
531 whereas deletions were separated by the DNA context of the deletion, being microhomologous,
532 repetitive, or none of the two. Microhomology was defined by whether the deleted sequence
533 was similar to the region immediately 3' of the breakpoint, as this indicates repair by
534 microhomology-mediated end joining. Repetitive deletions were defined by whether there is a
535 repeat of the indel at the 3' end of the breakpoint(23). We decided to not further compose indels
536 into signatures, for the ease of interpretation; the type of deletion has a clear relation to the
537 mechanism of repair, which is what we are investigating in this paper.

538

539 Finally, we used Signature Tools Lib to count structural variants: First, SVs were separated into
540 two groups, clustered and non-clustered based on the average distance between
541 rearrangements (the inter-rearrangement distance): Regions with at least 10 breakpoints having
542 at least 10 times smaller inter-rearrangement distance than the average across the genome of
543 that cancer, were considered clusters(61). SVs were further divided based on the type of
544 mutation: Deletions, tandem duplications, inversions and translocations. Deletions, tandem
545 duplications, and inversions were further divided by size, with intervals being 1-10Kb, 10-100Kb,
546 100Kb - 1Mb, 1-10Mb, and finally mutations with lengths above 10Mb. As with indels, we
547 decided not to decompose SV counts into rearrangement signatures, to ease interpretability and
548 because SV signatures are less established in the field. The per-sample exposure to each
549 summary statistic may be seen in **Additional file 1, Supp. Table 3** (log transformed and scaled
550 in **Additional file 1, Supp. Table 4;** see method below) and the suggested etiology (when
551 existing) may be seen in **Additional file 1, Supp. Table 5**(17,39).

552 # Statistical modeling

553 **Preparing subsets of data for modeling**

554 We separated the two data sets and divided the samples into their respective cohorts. We
555 stratified the data for each DDR gene; designating whether the patient had a biallelic hit,
556 monoallelic hit, or no mutation / only benign mutations. We excluded patients with inconclusive
557 variants (CADD phred >10,<25) within the gene. For modeling of cancers with biallelic hits in a
558 gene, we selected all DDR gene-cohort combinations where more than five patients had biallelic
559 hits in the same gene (n=194 LOF-sets). Likewise, we selected all cases with more than ten
560 samples with monoallelic hits (n = 341) (**Additional file 1, Supp. Table 6**). This setup means
561 that each sample may be included in developing several models, but occur only once in each
562 model as either mutated or non-mutated. Biallelic mutated samples were also included in the
563 models of monoallelic loss, as we consider biallelic loss a special case of monoallelic loss.
564

565 **Generating models using LASSO logistic regression**

566 For each LOF-set we calculated per-sample weights to counter the imbalance in the data, so
567 that the weight of each sample was one minus the proportion of samples with this mutational
568 status (either hit or non-hit/benign). We then used LASSO logistic regression from the R-
569 package “glmnet” (v4.0)(62) to select features with predictive power. For each of k cross-fold
570 validation sets (with k being the number of biallelic or monoallelic hit samples, respectively), we
571 ran a LASSO regression with a nested 5-fold cross-validation and the assigned weights (as
572 mentioned above to counter imbalance in the data), in order to produce a set of lambda values.
573 From the LASSO we selected the lambda value corresponding to the binomial deviance which
574 is one standard deviation away from the observed minimal binomial deviance of the regression,
575 in order to avoid overfitting (overfitting may occur if taking the lambda of the minimal binomial

576 deviance). In a few cases, the LASSO converged to larger sets of features than what is
577 justifiable by the low number of mutated samples, and so we limited the number of features to
578 one feature per ten mutated samples, rounded up and added one (e.g. 12 mutated samples
579 gives basis for maximum three features). For each of the k cross-folds we used the model to
580 predict the left-out data, and used these predictions for the evaluation of the model
581 performance.
582
583 Finally, we generated a model by including the entire data set. This was done in order to train
584 the best possible model in terms of selected features and coefficients, and this is the model
585 reported in the main figures, whereas the performance measures shown in the main figures are
586 derived from the k-fold cross-validation.
587
588 **Evaluating model predictive performance**
589 We used k-fold cross validation to get a measure of the predictive performance of each model.
590 Due to the strong imbalance in the data, we use the precision-recall area-under-the-curve (PR-
591 AUC) enrichment from the baseline (PR-AUC-E) as our statistic of performance, while also
592 reporting traditional AUROC scores. The baseline is defined as the PR-AUC obtained by a non-
593 informed model, which is equal to the true positive rate of the input data, e.g. 10 mutated out of
594 a 100 patients would have a baseline PR-AUC of 0.1. By subtracting the baseline from the PR-
595 AUC of each model we are able to compare the performance between models despite different
596 sample sizes.
597
598 **Model selection**
599 Across the 535 LOF-sets we observed a mean PR-AUC-E of 0.04 with a standard deviation of
600 0.099 (90%-quantile = 0.18) (**Additional file 1, Supp. Table 6**). We shortlisted models with PR-

601 AUC-E above two standard deviations above this (PR-AUC-E > 0.2) for further analysis
602 (**Additional file 1, Supp. Table 7**). To evaluate the significance of each model, we proceeded
603 to generate model-specific null-models as described below:
604
605 We generated a null-model for each case. This was done by permuting the mutation state of the
606 LOF-set and then running the LASSO logistic regression on the permuted data. We ran a
607 minimum of 10,000 permutations for each of the 535 cases, storing the PR-AUC-E for each
608 permutation. For LOF-sets where we observed fewer than five PR-AUC-E values more extreme
609 than the value from the true LOF-set, we ran an additional 10,000 models. This was continued
610 up to 30,000 permutations after which we stopped and concluded the p-value (**Additional file 1,**
611 **Supp. Table 6**). P-values were adjusted to consider the false discovery rate (FDR) of 535 tests,
612 and tests with FDR p-value above 0.05 were excluded from the model shortlist (**Additional file**
613 **1, Supp. Table 7**).
614
615 In conclusion, 48 models of 24 genes were shortlisted for further investigation. Of these, we
616 describe 43 in the main text and, more briefly, 5 in **Supplementary text** and **Supplementary**
617 **figure 1.** The inclusion into the main manuscript was done by collegial evaluation of interest to
618 the field and does not represent a definitive difference in quality of the models.

619 Declarations

620 Availability of Data and Materials

621 The analysis was based on two pre-existing data sets, that of the Pan-Cancer Analysis of Whole
622 Genomes and that of the Hartwig Medical Foundation. The public parts of the PCAWG data set

623 may be accessed at https://dcc.icgc.org/releases/PCAWG, whereas controlled files may be
624 accessed through gbGaP and daco, as instructed on this site https://docs.icgc.org/pcawg/data/.
625 The ICGC study ID of the project is EGAS00001001692.
626
627 The HMF data used in this project may be found by accession code DR-044 and can be
628 obtained upon request at the Hartwig Medical Foundation
629 (https://www.hartwigmedicalfoundation.nl/en).

630 Code availability

631 The code needed to reproduce the analysis is available at
632 https://github.com/SimonGrund/DDR_Predict , including data pre-processing, modeling of the
633 535 predictive models and modelling of ≥10,000 data permutation null models for each of the
634 535 models.

635 Competing interests

636 The authors declare no competing interests.

637 Funding

638 ERH and AS were funded by the Novo Nordisk Foundation (NNF15OC0016662), Cancer
639 Research UK (C23210/A7574), and the Danish National Research Foundation (Center
640 grant, DNRF115). JSP, SGS, GAP, and MHC were funded by the Independent Research

Fund Denmark | Medical Sciences (8021-00419B), a PhD stipend from Aarhus University, and a stipend from Research Foundation of Central Region Denmark (A2972).

Author contributions

The project was conceived and designed by JSP and SGS, with early contributions by JB. The data analysis was carried out by SGS under supervision of JSP with input from MHC, including data preparation, exploratory analysis, and statistical modeling. SGS generated all figures and tables. SGS, AS, GAP, ERH, and JSP interpreted the results. SGS, ERH and JSP wrote the manuscript. All authors accepted the manuscript.

Acknowledgements

We thank the Pan-Cancer Analysis of Whole Genomes (PCAWG), The International Cancer Genome Consortium (ICGC), and The Cancer Genome Atlas (TCGA) for access to whole cancer genomes. We would also like to thank the Hartwig Medical Foundation (HMF) and the Center for Personalized Cancer Treatment (CPCT) for creating and providing access to metastatic whole cancer genome data. We thank Jenny Gruhn and Amy V. Kaucher for valuable comments on the manuscript.

Reference list

1. Lindahl T. Instability and decay of the primary structure of DNA. Nature. 1993 Apr 22;362(6422):709–15.

2. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational Processes Molding the Genomes of 21 Breast Cancers [Internet]. Vol. 149, Cell. 2012. p. 979–93. Available from: http://dx.doi.org/10.1016/j.cell.2012.04.024

3. Volkova NV, Meier B, González-Huici V, Bertolini S, Gonzalez S, Vöhringer H, et al. Mutational signatures are jointly shaped by DNA damage and repair. Nat Commun. 2020 May 1;11(1):2169.

4. Bryant HE, Schultz N, Thomas HD, Parker KM, Flower D, Lopez E, et al. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase [Internet]. Vol. 447, Nature. 2007. p. 346–346. Available from: http://dx.doi.org/10.1038/nature05789

5. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N Engl J Med. 2015 Jun 25;372(26):2509–20.

6. Paculová H, Kramara J, Šimečková Š, Fedr R, Souček K, Hylse O, et al. BRCA1 or CDK12 loss sensitizes cells to CHK1 inhibitors. Tumour Biol. 2017 Oct;39(10):1010428317727479.

7. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014 Jan;42(Database issue):D980–5.

8. Davies H, Glodzik D, Morganella S, Yates LR, Staaf J, Zou X, et al. HRDetect is a predictor

681 of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat Med. 2017
682 Apr;23(4):517–25.

683 9. Lanni JS, Jacks T. Characterization of the p53-dependent postmitotic checkpoint following
684 spindle disruption. Mol Cell Biol. 1998 Feb;18(2):1055–64.

685 10. Gorgoulis VG, Vassiliou L-VF, Karakaidos P, Zacharatos P, Kotsinas A, Liloglou T, et al.
686 Activation of the DNA damage checkpoint and genomic instability in human precancerous
687 lesions. Nature. 2005 Apr 14;434(7035):907–13.

688 11. Umar A, Boyer JC, Thomas DC, Nguyen DC, Risinger JI, Boyd J, et al. Defective mismatch
689 repair in extracts of colorectal and endometrial cancer cell lines exhibiting microsatellite
690 instability. J Biol Chem. 1994 May 20;269(20):14367–70.

691 12. Edelmann W, Umar A, Yang K, Heyer J, Kucherlapati M, Lia M, et al. The DNA mismatch
692 repair genes Msh3 and Msh6 cooperate in intestinal tumor suppression. Cancer Res. 2000
693 Feb 15;60(4):803–7.

694 13. Popova T, Manié E, Boeva V, Battistella A, Goundiam O, Smith NK, et al. Ovarian Cancers
695 Harboring Inactivating Mutations in CDK12 Display a Distinct Genomic Instability Pattern
696 Characterized by Large Tandem Duplications. Cancer Res. 2016 Apr 1;76(7):1882–91.

697 14. Menghi F, Barthel FP, Yadav V, Tang M, Ji B, Tang Z, et al. The Tandem Duplicator
698 Phenotype Is a Prevalent Genome-Wide Cancer Configuration Driven by Distinct Gene
699 Mutations. Cancer Cell. 2018 Aug 13;34(2):197–210.e5.

700 15. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of
701 whole genomes. Nature. 2020 Feb;578(7793):82–93.

702 16. Priestley P, Baber J, Lolkema MP, Steeghs N, de Bruijn E, Shale C, et al. Pan-cancer

703 whole-genome analyses of metastatic solid tumours. Nature. 2019 Nov;575(7781):210–6.

704 17. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the
705 Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019 Jan 8;47(D1):D941–7.

706 18. Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations
707 hidden in cancer genomes. Curr Opin Genet Dev. 2014 Feb;24:52–60.

708 19. Pearl LH, Schierz AC, Ward SE, Al-Lazikani B, Pearl FMG. Therapeutic opportunities within
709 the DNA damage response. Nat Rev Cancer. 2015 Mar;15(3):166–80.

710 20. Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD, et al.
711 Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer
712 Genome Atlas. Cell Rep. 2018 Apr 3;23(1):239–54.e6.

713 21. Olivieri M, Cho T, Álvarez-Quilón A, Li K, Schellenberg MJ, Zimmermann M, et al. A
714 Genetic Map of the Response to DNA Damage in Human Cells. Cell. 2020 Jul
715 23;182(2):481–96.e21.

716 22. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the
717 deleteriousness of variants throughout the human genome [Internet]. Vol. 47, Nucleic Acids
718 Research. 2019. p. D886–94. Available from: http://dx.doi.org/10.1093/nar/gky1016

719 23. Degasperi A, Amarante TD, Czarnecki J, Shooter S, Zou X, Glodzik D, et al. A practical
720 framework and online tool for mutational signature analyses show inter-tissue variation and
721 driver dependencies. Nat Cancer. 2020 Feb;1(2):249–63.

722 24. Nguyen L, W M Martens J, Van Hoeck A, Cuppen E. Pan-cancer landscape of homologous
723 recombination deficiency. Nat Commun. 2020 Nov 4;11(1):5584.

724 25. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In:
725 Proceedings of the 23rd international conference on Machine learning. New York, NY,
726 USA: Association for Computing Machinery; 2006. p. 233–40. (ICML ’06).

727 26. Ceccaldi R, Liu JC, Amunugama R, Hajdu I, Primack B, Petalcorin MIR, et al. Homologous-
728 recombination-deficient tumours are dependent on Polθ-mediated repair. Nature. 2015 Feb
729 12;518(7538):258–62.

730 27. Hanel W, Moll UM. Links between mutant p53 and genomic instability. J Cell Biochem.
731 2012 Feb;113(2):433–9.

732 28. Graham LS, Montgomery B, Cheng HH, Yu EY, Nelson PS, Pritchard C, et al. Mismatch
733 repair deficiency in metastatic prostate cancer: Response to PD-1 blockade and standard
734 therapies. PLoS One. 2020 May 26;15(5):e0233260.

735 29. Blazek D, Kohoutek J, Bartholomeeusen K, Johansen E, Hulinkova P, Luo Z, et al. The
736 Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA
737 damage response genes. Genes Dev. 2011 Oct 15;25(20):2158–72.

738 30. Marqués F, Moreau JL, Peaucellier G, Lozano JC, Schatt P, Picard A, et al. A new
739 subfamily of high molecular mass CDC2-related kinases with PITAI/VRE motifs. Biochem
740 Biophys Res Commun. 2000 Dec 29;279(3):832–7.

741 31. Li X, Chatterjee N, Spirohn K, Boutros M, Bohmann D. Cdk12 Is A Gene-Selective RNA
742 Polymerase II Kinase That Regulates a Subset of the Transcriptome, Including Nrf2 Target
743 Genes. Sci Rep. 2016 Feb 25;6:21455.

744 32. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of
745 somatic structural variation in human cancer genomes. Nature. 2020 Feb;578(7793):112–

21.

33. Wu Y-M, Cieślik M, Lonigro RJ, Vats P, Reimers MA, Cao X, et al. Inactivation of CDK12 Delineates a Distinct Immunogenic Class of Advanced Prostate Cancer [Internet]. Vol. 173, Cell. 2018. p. 1770–82.e14. Available from: http://dx.doi.org/10.1016/j.cell.2018.04.034

34. Rescigno P, Gurel B, Pereira R, Crespo M, Rekowski J, Rediti M, et al. Characterizing CDK12-Mutated Prostate Cancers. Clin Cancer Res. 2021 Jan 15;27(2):566–74.

35. Sumanasuriya S, De Bono J. Treatment of Advanced Prostate Cancer—A Review of Current Therapies and Future Promise. Cold Spring Harb Perspect Med [Internet]. 2018 Jun 1;8(6). Available from: http://perspectivesinmedicine.cshlp.org/content/8/6/a030635.abstract

36. Bajrami I, Frankum JR, Konde A, Miller RE, Rehman FL, Brough R, et al. Genome-wide profiling of genetic synthetic lethality identifies CDK12 as a novel determinant of PARP1/2 inhibitor sensitivity. Cancer Res. 2014 Jan 1;74(1):287–97.

37. Joshi PM, Sutor SL, Huntoon CJ, Karnitz LM. Ovarian cancer-associated mutations disable catalytic activity of CDK12, a kinase that promotes homologous recombination repair and resistance to cisplatin and poly(ADP-ribose) polymerase inhibitors. J Biol Chem. 2014 Mar 28;289(13):9247–53.

38. Mukherjee J, Johannessen T-C, Ohba S, Chow TT, Jones L, Pandita A, et al. Mutant IDH1 Cooperates with ATRX Loss to Drive the Alternative Lengthening of Telomere Phenotype in Glioma. Cancer Res. 2018 Jun 1;78(11):2966–77.

39. 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;578(7793):94–101.

768 40. Singh VK, Rastogi A, Hu X, Wang Y, De S. Mutational signature SBS8 predominantly
769 arises due to late replication errors in cancer. Commun Biol. 2020 Aug 3;3(1):421.

770 41. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like
771 mutational processes in human somatic cells. Nat Genet. 2015 Dec;47(12):1402–7.

772 42. Alexandrov LB, Australian Pancreatic Cancer Genome Initiative, Nik-Zainal S, Wedge DC,
773 Samuel A J, Behjati S, et al. Signatures of mutational processes in human cancer [Internet].
774 Vol. 500, Nature. 2013. p. 415–21. Available from: http://dx.doi.org/10.1038/nature12477

775 43. Nik-Zainal S, Kucab JE, Morganella S, Glodzik D, Alexandrov LB, Arlt VM, et al. The
776 genome as a record of environmental exposure [Internet]. Mutagenesis. 2015. p. gev073.
777 Available from: http://dx.doi.org/10.1093/mutage/gev073

778 44. Sauter JL, Graham RP, Larsen BT, Jenkins SM, Roden AC, Boland JM. SMARCA4-
779 deficient thoracic sarcoma: a distinctive clinicopathological entity with undifferentiated
780 rhabdoid morphology and aggressive behavior. Mod Pathol. 2017 Oct;30(10):1422–32.

781 45. Rekhtman N, Montecalvo J, Chang JC, Alex D, Ptashkin RN, Ai N, et al. SMARCA4-
782 Deficient Thoracic Sarcomatoid Tumors Represent Primarily Smoking-Related
783 Undifferentiated Carcinomas Rather Than Primary Thoracic Sarcomas. J Thorac Oncol.
784 2020 Feb;15(2):231–47.

785 46. Dagogo-Jack I, Schrock AB, Kem M, Jessop N, Lee J, Ali SM, et al. Clinicopathologic
786 Characteristics of BRG1-Deficient NSCLC. J Thorac Oncol. 2020 May;15(5):766–76.

787 47. Concepcion CP, Ma S, LaFave LM, Bhutkar A, Liu M, DeAngelo LP, et al. SMARCA4
788 inactivation promotes lineage-specific transformation and early metastatic features in the
789 lung. Cancer Discov [Internet]. 2021 Sep 24; Available from: http://dx.doi.org/10.1158/2159-

790 8290.CD-21-0248

791 48. Liu L, Dilworth D, Gao L, Monzon J, Summers A, Lassam N, et al. Mutation of the CDKN2A
792 5'UTR creates an aberrant initiation codon and predisposes to melanoma. Nat Genet.
793 1999;21(1):128–32.

794 49. Amos CI, Wang L-E, Lee JE, Gershenwald JE, Chen WV, Fang S, et al. Genome-wide
795 association study identifies novel loci predisposing to cutaneous melanoma. Hum Mol
796 Genet. 2011 Dec 15;20(24):5012–23.

797 50. Cubillos-Rojas M, Amair-Pinedo F, Peiró-Jordán R, Bartrons R, Ventura F, Rosa JL. The
798 E3 ubiquitin protein ligase HERC2 modulates the activity of tumor protein p53 by regulating
799 its oligomerization. J Biol Chem. 2014 May 23;289(21):14782–95.

800 51. Li J, Yen C, Liaw D, Podsypanina K, Bose S, Wang SI, et al. PTEN, a putative protein
801 tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. Science.
802 1997 Mar 28;275(5308):1943–7.

803 52. Liaw D, Marsh DJ, Li J, Dahia PL, Wang SI, Zheng Z, et al. Germline mutations of the
804 PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. Nat
805 Genet. 1997 May;16(1):64–7.

806 53. Fda. FDA approves niraparib for HRD-positive advanced ovarian cancer. 2019;

807 54. Boland CR, Goel A. Microsatellite instability in colorectal cancer. Gastroenterology. 2010
808 Jun;138(6):2073–87.e3.

809 55. Umar A, Boland CR, Terdiman JP, Syngal S, Chapelle A de la, Rüschoff J, et al. Revised
810 Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and
811 microsatellite instability. J Natl Cancer Inst. 2004;96(4):261–8.

812 56. Abida W, Cheng ML, Armenia J, Middha S, Autio KA, Vargas HA, et al. Analysis of the
813 Prevalence of Microsatellite Instability in Prostate Cancer and Response to Immune
814 Checkpoint Blockade. JAMA Oncol. 2019 Apr 1;5(4):471–8.

815 57. Antonarakis ES, Shaukat F, Isaacsson Velho P, Kaur H, Shenderov E, Pardoll DM, et al.
816 Clinical Features and Therapeutic Outcomes in Men with Advanced Prostate Cancer and
817 DNA Mismatch Repair Gene Mutations. Eur Urol. 2019 Mar;75(3):378–82.

818 58. Xue Y, Meehan B, Fu Z, Wang XQD, Fiset PO, Rieker R, et al. SMARCA4 loss is synthetic
819 lethal with CDK4/6 inhibition in non-small cell lung cancer. Nat Commun. 2019 Feb
820 4;10(1):557.

821 59. Malcikova J, Smardova J, Rocnova L, Tichy B, Kuglik P, Vranova V, et al. Monoallelic and
822 biallelic inactivation of TP53 gene in chronic lymphocytic leukemia: selection, impact on
823 survival, and response to DNA damage. Blood. 2009 Dec 17;114(26):5307–14.

824 60. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The
825 mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020
826 May;581(7809):434–43.

827 61. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of
828 somatic mutations in 560 breast cancer whole-genome sequences. Nature. 2016 Jun
829 2;534(7605):47–54.

830 62. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via
831 coordinate descent. J Stat Softw [Internet]. 2010; Available from:
832 https://www.ncbi.nlm.nih.gov/pmc/articles/pmc2929880/

833 63. Trizzino M, Barbieri E, Petracovici A, Wu S, Welsh SA, Owens TA, et al. The Tumor

834 Suppressor ARID1A Controls Global Transcription via Pausing of RNA Polymerase II. Cell
835 Rep. 2018 Jun 26;23(13):3933–45.

836 64. Pilarski R, Carlo M, Cebulla C, Abdel-Rahman M. BAP1 tumor predisposition syndrome.
837 2020; Available from: https://europepmc.org/books/nbk390611/

838 65. Pfeifer GP, You Y-H, Besaratinia A. Mutations induced by ultraviolet light. Mutat Res. 2005
839 Apr 1;571(1-2):19–31.

840 66. Howard BD, Tessman I. Identification of the altered bases in mutated single-stranded DNA:
841 III. Mutagenesis by ultraviolet light. J Mol Biol. 1964 Aug 1;9(2):372–5.

842 67. Drost J, van Boxtel R, Blokzijl F, Mizutani T, Sasaki N, Sasselli V, et al. Use of CRISPR-
843 modified human stem cell organoids to study the origin of mutational signatures in cancer.
844 Science. 2017 Oct 13;358(6360):234–8.

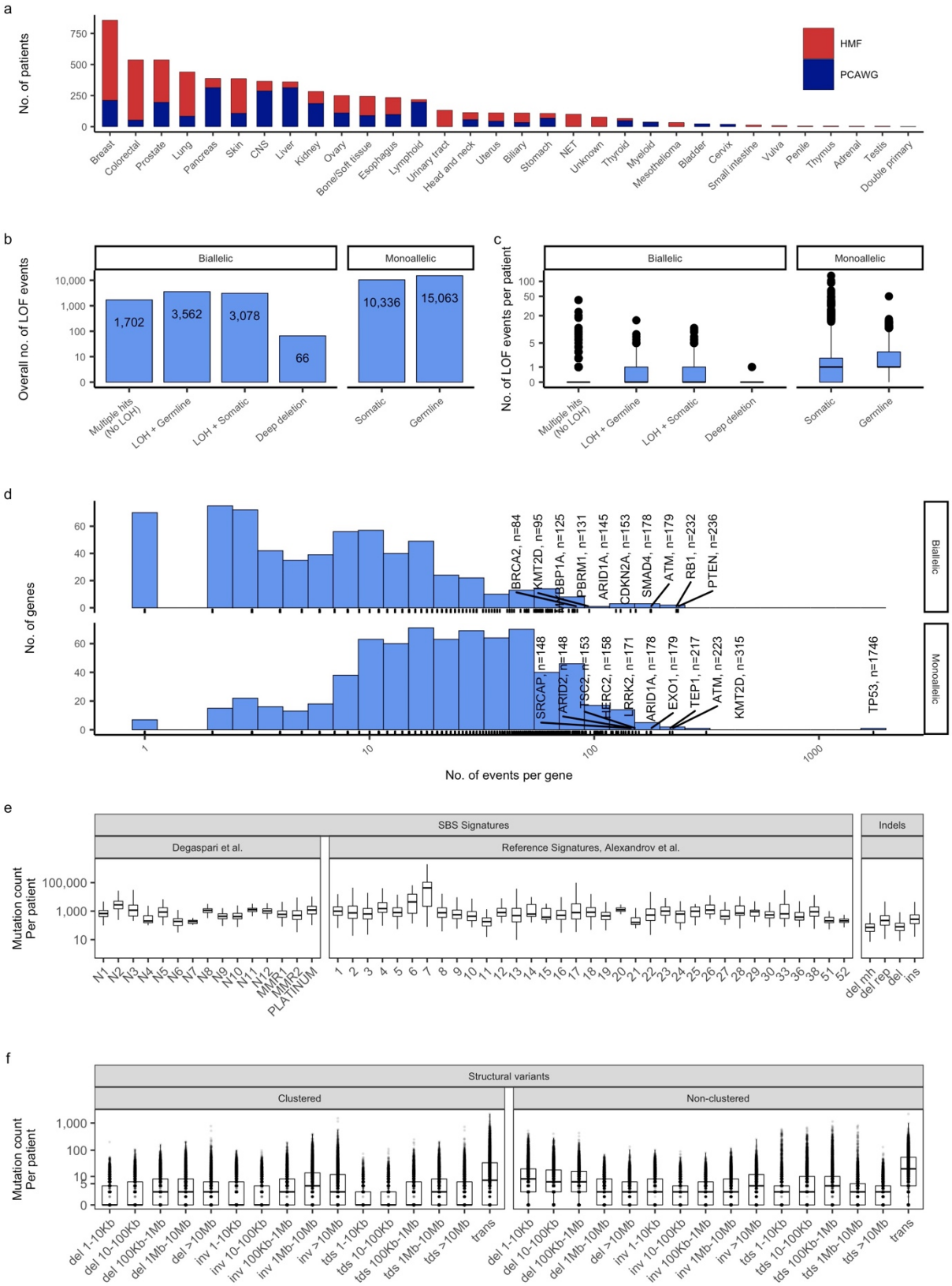
845 68. Chandrasekharappa SC, Guru SC, Manickam P, Olufemi SE, Collins FS, Emmert-Buck
846 MR, et al. Positional cloning of the gene for multiple endocrine neoplasia-type 1. Science.
847 1997 Apr 18;276(5311):404–7.

848 69. Heppner C, Kester MB, Agarwal SK, Debelenko LV, Emmert-Buck MR, Guru SC, et al.
849 Somatic mutation of the MEN1 gene in parathyroid tumours. Nat Genet. 1997
850 Aug;16(4):375–8.

851

852 # Figures and legends

Figure 1


a
No. of patients
HMF
PCAWG
Breast
Colorectal
Prostate
Lung
Pancreas
Skin
CNS
Liver
Kidney
Ovary
Bone/Soft tissue
Esophagus
Lymphoid
Urinary tract
Head and neck
Uterus
Biliary
Stomach
NET
Unknown
Thyroid
Myeloid
Mesothelioma
Bladder
Cervix
Small intestine
Vulva
Penile
Thymus
Adrenal
Testis
Double primary
b
Overall no. of LOF events
Biallelic
Monoallelic
1,702
3,562
3,078
66
10,336
15,063
Multiple hits (No LOH)
LOH + Germline
LOH + Somatic
Deep deletion
Somatic
Germline
c
No. of LOF events per patient
d
No. of genes
BRCA2, n=84
KMT2D, n=95
BBP1A, n=125
PBRM1, n=131
ARID1A, n=145
CDKN2A, n=153
SMAD4, n=178
ATM, n=179
RB1, n=232
PTEN, n=236
SRCAP, n=148
ARID2, n=148
TSC2, n=153
ERC2, n=158
LRRK2, n=171
ARID1A, n=178
EXO1, n=179
TEP1, n=217
ATM, n=223
KMT2D, n=315
TP53, n=1746
No. of events per gene
e
SBS Signatures
Indels
Degaspari et al.
Reference Signatures, Alexandrov et al.
Mutation count Per patient
MMR1
MMR2
PLATINUM
del mh
del rep
del
ins
f
Structural variants
Clustered
Non-clustered
del 1-10Kb
del 10-100Kb
del 100Kb-1Mb
del 1Mb-10Mb
del >10Mb
inv 1-10Kb
inv 10-100Kb
inv 100Kb-1Mb
inv 1Mb-10Mb
inv >10Mb
tds 1-10Kb
tds 10-100Kb
tds 100Kb-1Mb
tds 1Mb-10Mb
tds >10Mb
trans

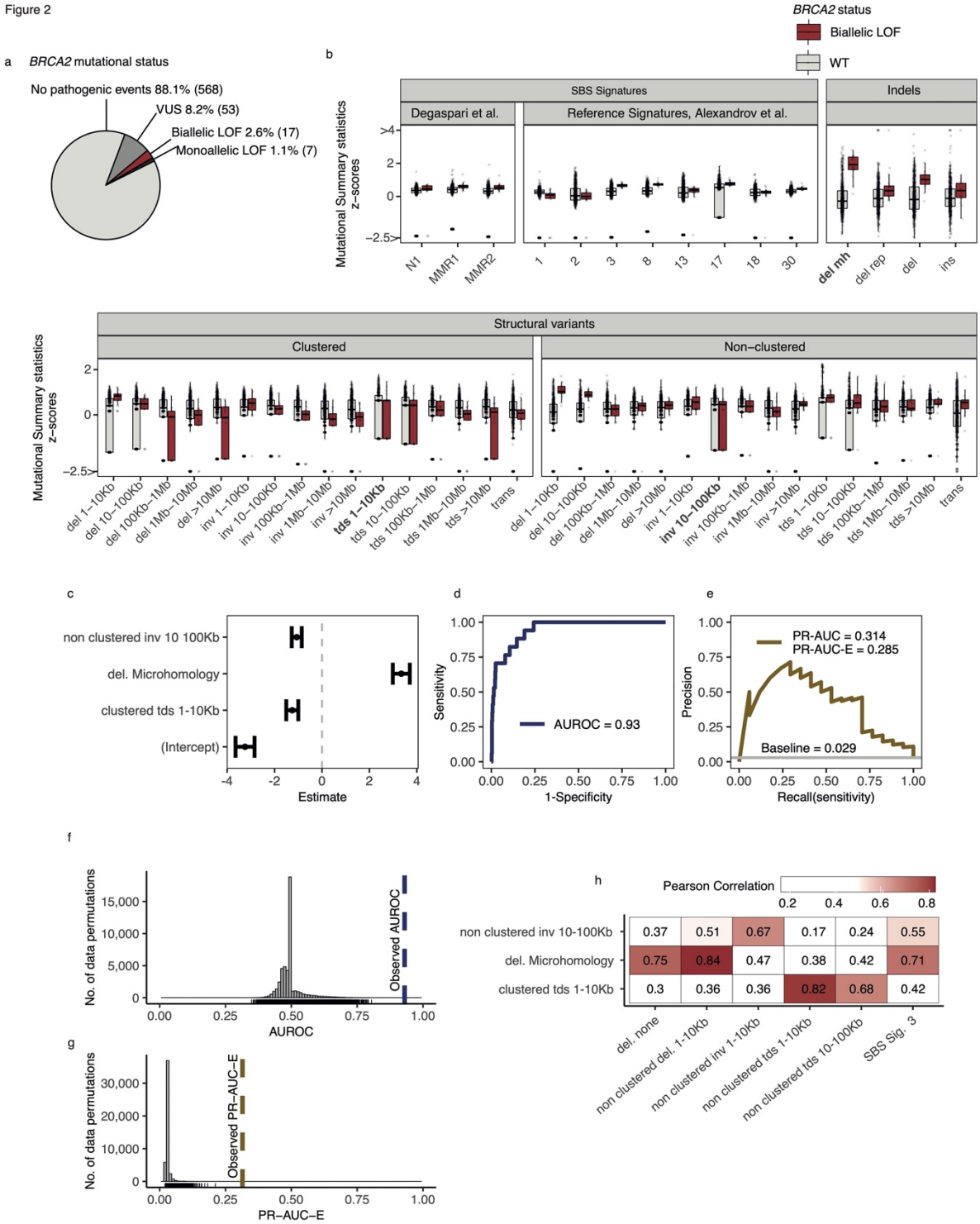
853

854 **Figure 1: Cancer-types, DDR gene deficiencies, and mutational summary statistics**

855 **a** Cohort sizes for the 32 cancer types comprising the 6,065 whole cancer genomes collected
856 from the Hartwig Medical Foundation (HMF; n=3,497) and the PanCancer Analysis of Whole
857 Genomes (PCAWG; n=2,568). **b** mono- and biallelic loss-of-function (LOF) events were
858 annotated across 736 DDR genes based on both pathogenic variants and copy-number-losses
859 (loss of heterozygosity; LOH), overall and (**c**) per patient (**d**) with varying no. of LOF events per
860 DDR gene (x-axis; logarithmic). **e** Whole genome mutational patterns were represented as
861 summary statistics and used as input features for the predictive models of DDR gene deficiency.
862 Concretely, each patient was annotated with the number of single-base-substitutions (SBS) that
863 are accounted to each SBS signature (Alexandrov et al.(18); Degaspari et al.(23)), number of
864 indels divided by context (mh = microhomology; rep = repetitive), and (**f**) number of structural
865 variants divided by clusterness, size, and type (del = deletion; inv = inversion; tds = tandem
866 duplication; trans = translocation)

867

868

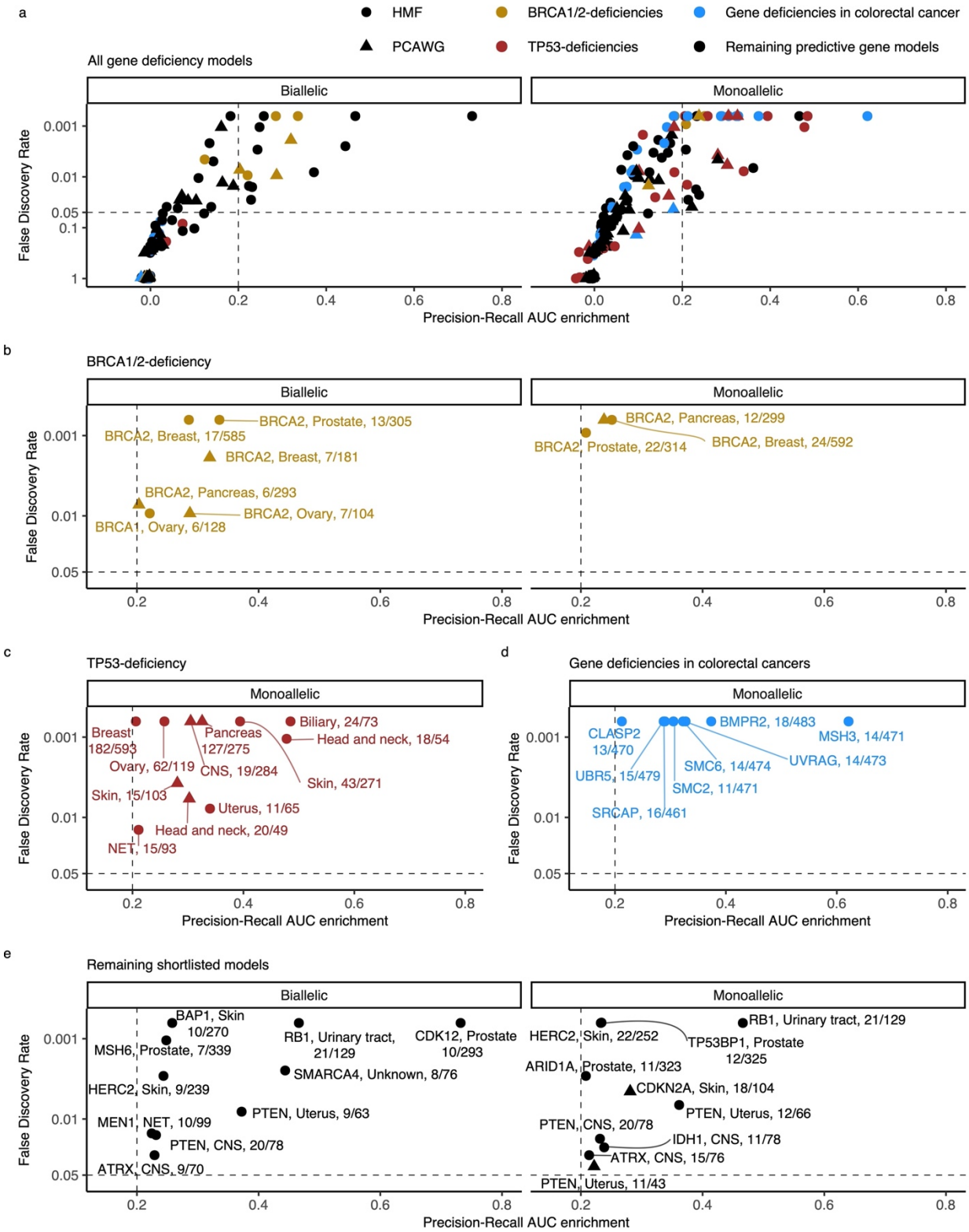


869

870 **Figure 2: Predictive modelling of BRCA2 deficiencies in the HMF breast cancers**

871 **a** Mutational status of *BRCA2* across 645 HMF breast cancer patients. **b** Mutational summary
872 statistics for the HMF breast cancer patients divided by biallelic *BRCA2* loss-of-function (LOF;
873 red) and *BRCA2* wild type (WT; grey) (selected predictive features in bold). **c** Predictive features
874 and their coefficients for model of biallelic *BRCA2* loss with predictive performance measured in
875 (**d**) AUROC and (**e**) PR-AUC (PR-AUC-E=PR-AUC - baseline=0.29; **Methods**). **f** Distributions of
876 AUROC and (**g**) PR-AUC-E values obtained from 30,000 random data permutations compared
877 to observed values (punctuated lines). **h** Correlation between selected predictive features
878 (horisontal) and other highly correlated (Pearson corr. >0.65) mutational features (vertical).
879

Figure 3


a
HMF
PCAWG
BRCA1/2-deficiencies
TP53-deficiencies
Gene deficiencies in colorectal cancer
Remaining predictive gene models
All gene deficiency models
Biallelic
Monoallelic
False Discovery Rate
Precision-Recall AUC enrichment
b
BRCA1/2-deficiency
BRCA2, Prostate, 13/305
BRCA2, Breast, 17/585
BRCA2, Breast, 7/181
BRCA2, Pancreas, 6/293
BRCA2, Ovary, 7/104
BRCA1, Ovary, 6/128
BRCA2, Pancreas, 12/299
BRCA2, Breast, 24/592
BRCA2, Prostate, 22/314
c
TP53-deficiency
Breast 182/593
Pancreas 127/275
Biliary, 24/73
Head and neck, 18/54
Ovary, 62/119
CNS, 19/284
Skin, 43/271
Skin, 15/103
Uterus, 11/65
Head and neck, 20/49
NET, 15/93
d
Gene deficiencies in colorectal cancers
CLASP2 13/470
BMPR2, 18/483
MSH3, 14/471
UVRAG, 14/473
SMC6, 14/474
UBR5, 15/479
SMC2, 11/471
SRCAP, 16/461
e
Remaining shortlisted models
BAP1, Skin 10/270
MSH6, Prostate, 7/339
RB1, Urinary tract, 21/129
CDK12, Prostate 10/293
SMARCA4, Unknown, 8/76
HERC2, Skin, 9/239
PTEN, Uterus, 9/63
MEN1, NET, 10/99
PTEN, CNS, 20/78
ATRX, CNS, 9/70
HERC2, Skin, 22/252
RB1, Urinary tract, 21/129
TP53BP1, Prostate 12/325
ARID1A, Prostate, 11/323
CDKN2A, Skin, 18/104
PTEN, Uterus, 12/66
PTEN, CNS, 20/78
IDH1, CNS, 11/78
ATRX, CNS, 15/76
PTEN, Uterus, 11/43

880

881 **Figure 3: Predictive models of DDR gene deficiencies**

882 **a** The predictive performance (PR-AUC-E; x-axis) and significance (FDR; logarithmic y-axis) of
883 the 535 predictive models (one model per gene with more than five biallelic or more than ten
884 patients either mono- or biallelic mutated in either HMF or PCAWG in any one cancer type;
885 **Methods**). Significance (q-value representing FDR) evaluated by counting equally or more-
886 extreme PR-AUC-E values across >10,000 permuted data sets and applying Benjamini-
887 Hochberg FDR control. Models with FDR below 0.05 and PR-AUC-E above 0.2 are shortlisted
888 (**Methods**). **b** Shortlisted predictive models of deficiency of *BRCA1* or *BRCA2*; (**c**) *TP53*
889 monoallelic predictive models; (**d**) monoallelic gene deficiency models across colorectal cancer
890 patients; (**e**) and remaining gene deficiency models not contained in the other sub-groups.

891

892

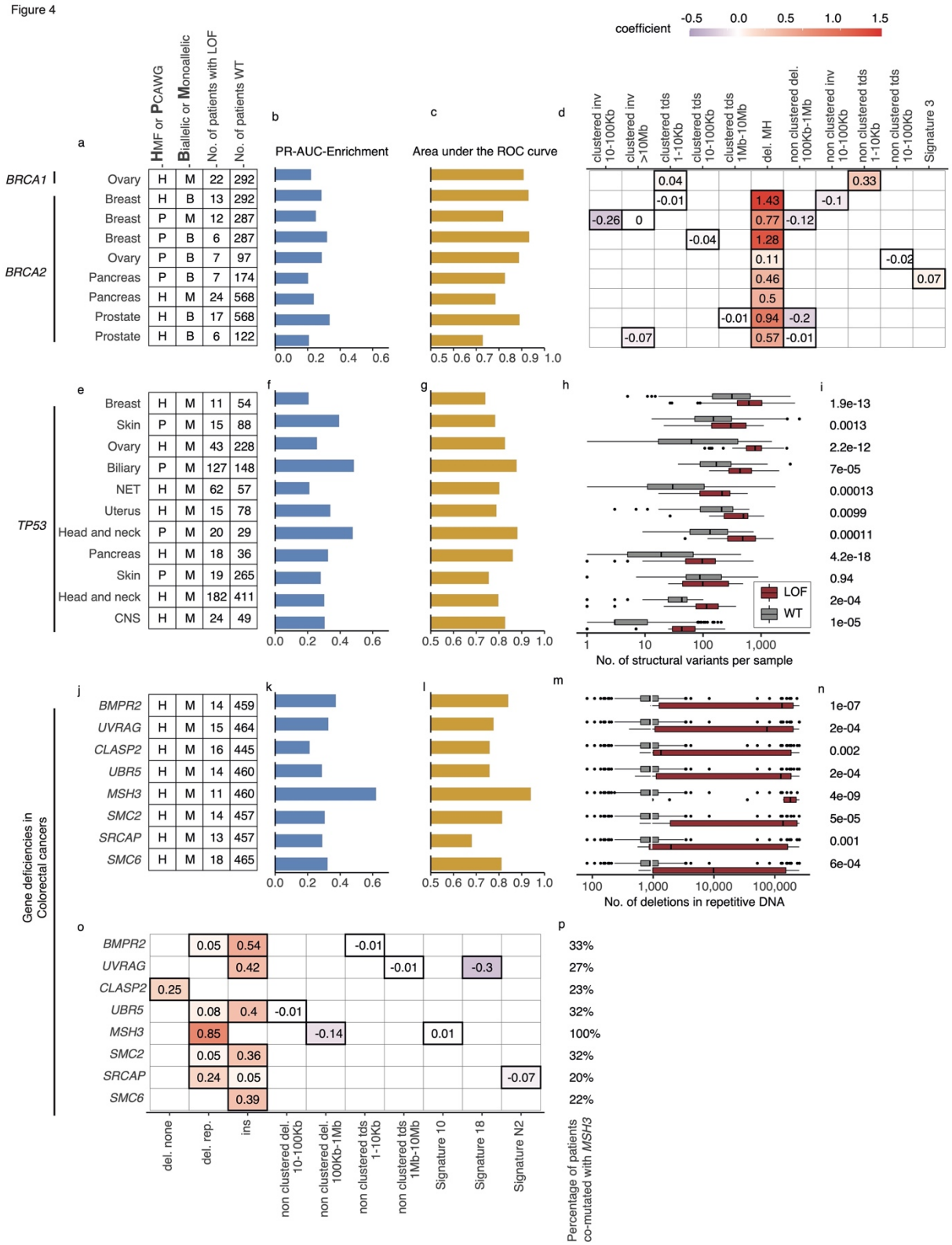


Figure 4: Predictive models with anticipated etiology or origin

895 **a** Overview of predictive models for *BRCA1*-d and *BRCA2*-d, showing data source, type of
896 model, and LOF-set statistics *(***b**) PR-AUC-E, (**c**) AUROC, (**d**) and the predictive features and
897 their coefficient for individual models. **e,f,g** Overview of predictive models of *TP53*-d (as in a-c).
898 **h** For each cohort the number of structural variants (x-axis; logarithmic) for *TP53* LOF patients
899 (red) versus *TP53* wild type patients (grey) and (**i**) the significance of their difference (two-sided
900 Wilcoxon rank-sum test). **j,k,l** Predictive models of gene deficiencies in colorectal cancers (as
901 in a-c). **m** Number of deletions in repetitive DNA (as in h) and (**n**) its significance (as in i)**. o** The
902 predictive features of each model (as in d) and (**p**) the percentage of patients that are co-
903 mutated with *MSH3*.
904

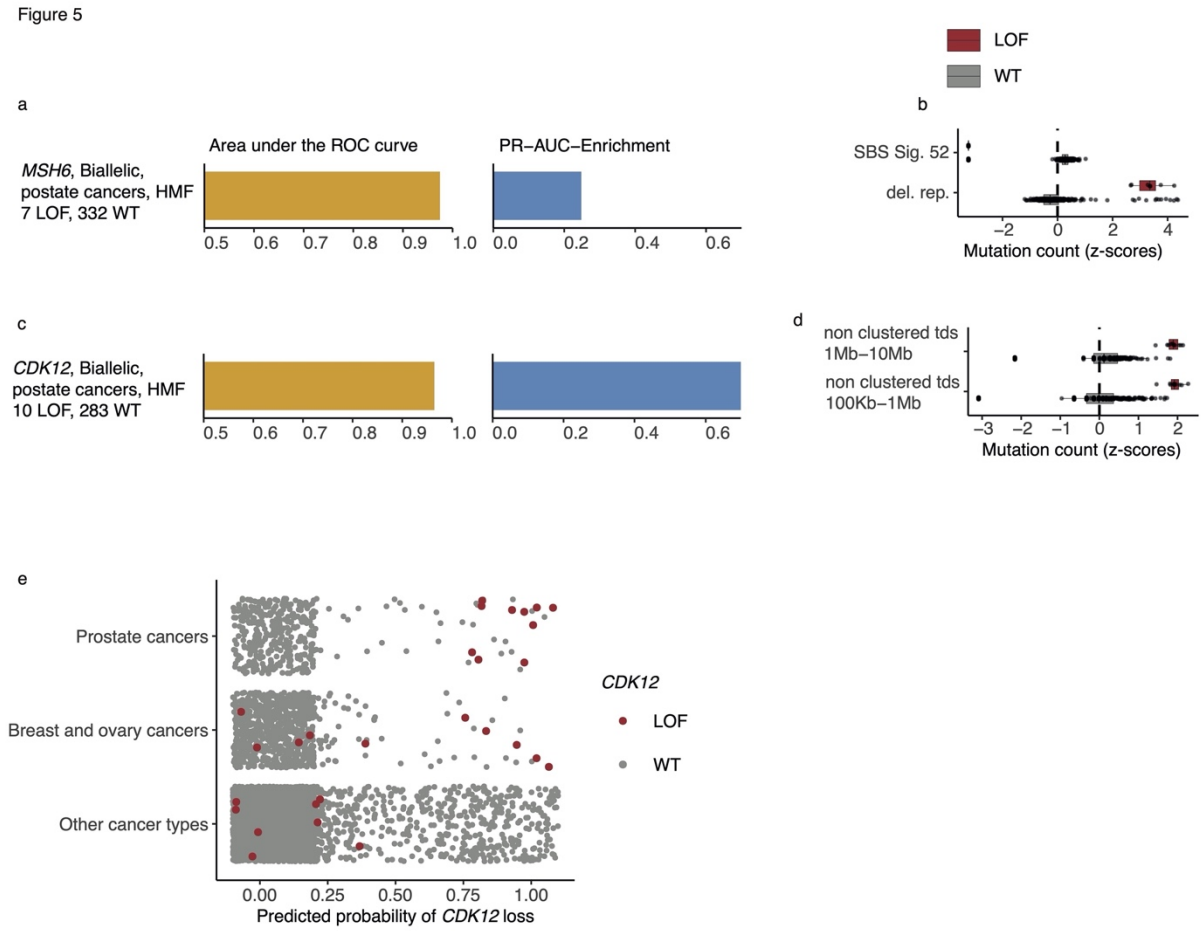
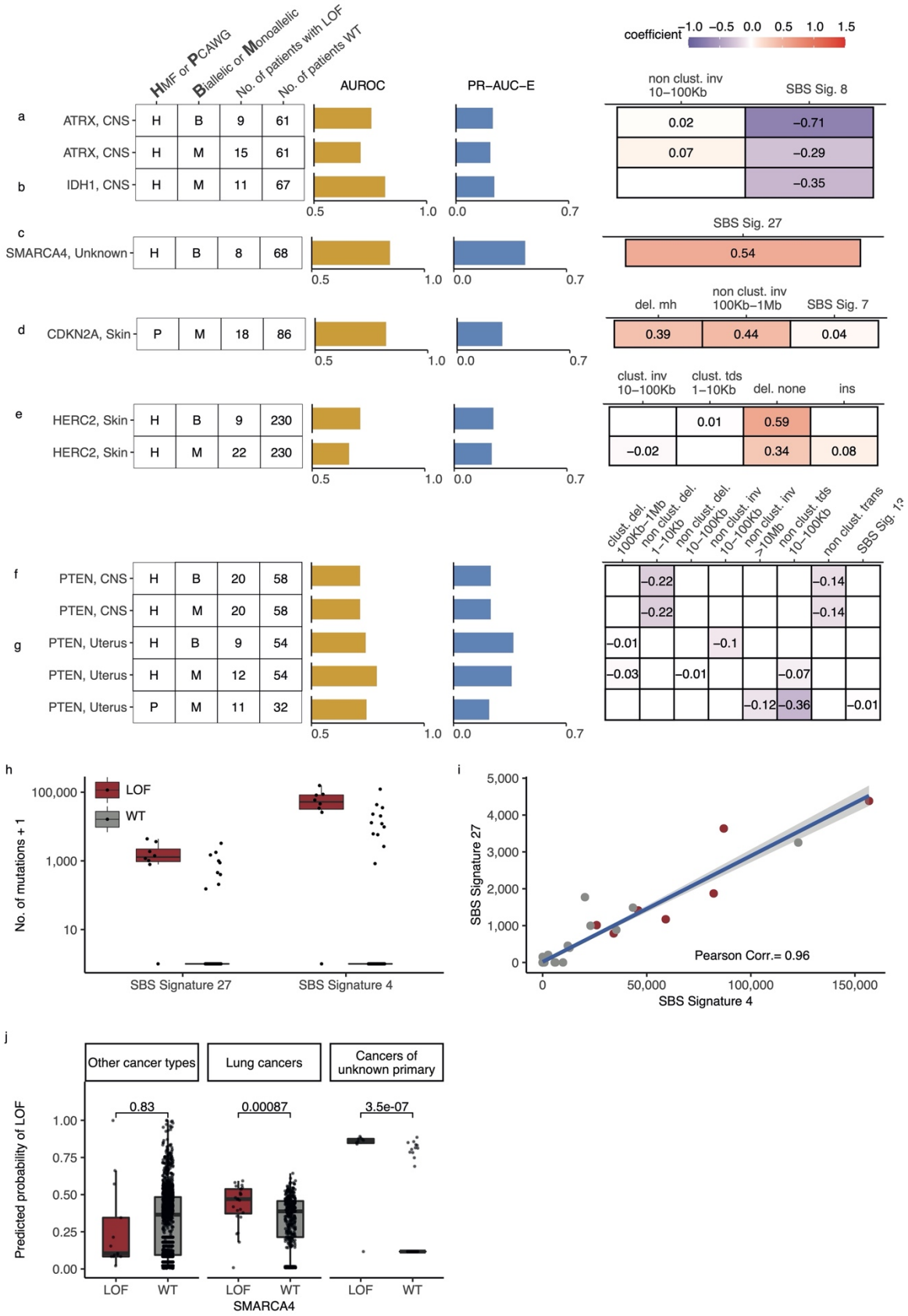


Figure 5: *CDK12* mutated prostate tumours are predicted by tandem duplications

a Biallelic predictive model *MSH6*-d in HMF prostate cancer patients and its PR-AUC-E and AUROC. **b** Boxplots of mutation counts between patients that are *MSH6*-d loss-of-function (red) and *MSH6* wild-type (grey), (mutation counts are normalised and log-transformed; **Methods). c** Biallelic predictive model for *CDK12*-d with performance measures (as in a). **b** Boxplots of mutation counts for *CDK12* deficient and wildtype patients (as in b). **e** *CDK12*-d predictive performance for different cancer types. Predicted probability of *CDK12*-d (x-axis) for patients with *CDK12* LOF (red) and *CDK12* wild type (grey) are shown for prostate cancer, breast and ovary cancer, and all other cancer types (y-axis).

Figure 6


HMF or PCAWG
Biallelic or Monoallelic
No. of patients with LOF
No. of patients WT
AUROC
PR–AUC–E
coefficient
a ATRX, CNS H B 9 61
ATRX, CNS H M 15 61
b IDH1, CNS H M 11 67
c SMARCA4, Unknown H B 8 68
d CDKN2A, Skin P M 18 86
e HERC2, Skin H B 9 230
HERC2, Skin H M 22 230
f PTEN, CNS H B 20 58
PTEN, CNS H M 20 58
g PTEN, Uterus H B 9 54
PTEN, Uterus H M 12 54
PTEN, Uterus P M 11 32
non clust. inv 10–100Kb
SBS Sig. 8
SBS Sig. 27
del. mh
non clust. inv 100Kb–1Mb
SBS Sig. 7
clust. inv 10–100Kb
clust. tds 1–10Kb
del. none
ins
clust. del. 100Kb–1Mb
non clust. del. 1–10Kb
non clust. del. 10–100Kb
non clust. inv 10–100Kb
non clust. inv >10Mb
non clust. tds 10–100Kb
non clust. trans
SBS Sig. 13
h
LOF
WT
No. of mutations + 1
SBS Signature 27
SBS Signature 4
i
SBS Signature 27
SBS Signature 4
Pearson Corr.= 0.96
j
Other cancer types
Lung cancers
Cancers of unknown primary
Predicted probability of LOF
0.83
0.00087
3.5e-07
SMARCA4

915

916 **Figure 6: Novel predictive models of DDR gene deficiencies**

917 **a** Predictive model of *ATRX*-d and its PR-AUC-E, AUROC, and selected features and their
918 coefficients. Same information for predictive models of (**b)** *IDH1*-d, (**c**) *SMARCA4*-d, (**d**)
919 *CDKN2A*-d, (**e**) *HERC2*-d, and (**f**) *PTEN*-d in CNS cancers and (**g**) uterine cancers.
920 **h** Number of SBS sig. 27 and SBS sig. 4 (y-axis; logarithmic) among patients of unknown
921 primary with *SMARCA4* biallelic LOF (red) or wild-type (grey). **i** Pearson correlation between the
922 per-patient number of SBS sig 27 (y-axis) and SBS sig 4 (x-axis; logarithmic) mutations, with an
923 overlaid linear model (blue) and its 95% confidence interval (grey). **j** Using a model trained to
924 predict *SMARCA4* biallelic LOF in HMF cancers of unknown primary, we evaluate the predictive
925 power across cohorts (one-tailed wilcoxon rank-sum test) displaying significant cohorts
926 separately (colours as in **a**).

Additional files

Additional file 1: Supplementary tables 1,2 and 4-9; Microsoft Excel (xlsx)

Supp. Table 1: All included samples and their primary tumour locations

Supp. Table 2: 736 DDR genes, hg19 coordinates and the number of pathogenic events across 6,065 cancer genomes

Supp. Table 5: Proposed Etiologies of base substitution signatures

Supp. Table 6: All models (n=535)

Supp. Table 7: Shortlisted models (n=48)

Supp. Table 8: Pathogenic events in each of the 535 LOF-sets

Supp. Table 9: Correlation between features in shortlisted models

Additional file 2.1: Supp. Table 3: All SBS signature contributions, indels counts, and SV counts, tab-separated values (.tsv), may be opened in Microsoft Excel per sample

Additional file 2.2: Supp. Table 4: All SBS signature contributions, indels counts, and SV counts, per sample, log-transformed and scaled to z-scores, tab-separated values (.tsv), may be opened in Microsoft Excel per sample

Additional file 3: Supplementary text; pdf (.pdf)

Supplementary Text: Significant models that are not described in the main text

948 Additional file 4: Supplementary figures; pdf (.pdf)

949 Supp. Fig. 1: Shortlisted predictive models of DDR gene deficiencies not included in the main

950 manuscript

951 Supp. Fig. 2: CDKN2A-d in PCAWG and HMF skin cancers, HERC2-d in HMF skin cancers,

952 and ARID1A-d in PCAWG prostate cancers

953