# HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures

Helen Davies, Dominik Glodzik, Sandro Morganella, Lucy R Yates, Johan Staaf, Xueqing Zou, Manasa Ramakrishna, Sancha Martin, Sandrine Boyault, Anieta M Sieuwerts, et al.

▶ **To cite this version:**

Helen Davies, Dominik Glodzik, Sandro Morganella, Lucy R Yates, Johan Staaf, et al.. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nature Medicine, 2017, 23 (4), pp.517 - 525. 10.1038/nm.4292 . hal-01525050

**HAL Id: hal-01525050**
**https://inria.hal.science/hal-01525050**

Submitted on 12 Jul 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# HRDetect is a predictor of *BRCA1* and *BRCA2* deficiency based on mutational signatures

Helen Davies[1,32], Dominik Glodzik[1,32], Sandro Morganella[1], Lucy R Yates[1,2], Johan Staaf[3], Xueqing Zou[1], Manasa Ramakrishna[1,4], Sancha Martin[1], Sandrine Boyault[5], Anieta M Sieuwerts[6], Peter T Simpson[7], Tari A King[8], Keiran Raine[1], Jorunn E Eyfjord[9], Gu Kong[10], Åke Borg[3], Ewan Birney[11], Hendrik G Stunnenberg[12], Marc J van de Vijver[13], Anne-Lise Børresen-Dale[14,15], John W M Martens[6], Paul N Span[16,17], Sunil R Lakhani[7,18], Anne Vincent-Salomon[19,20], Christos Sotiriou[21], Andrew Tutt[22,23], Alastair M Thompson[24], Steven Van Laere[25,26], Andrea L Richardson[27,28], Alain Viari[29,30], Peter J Campbell[1], Michael R Stratton[1] & Serena Nik-Zainal[1,31]

**Approximately 1–5% of breast cancers are attributed to inherited mutations in *BRCA1* or *BRCA2* and are selectively sensitive to poly(ADP-ribose) polymerase (PARP) inhibitors. In other cancer types, germline and/or somatic mutations in *BRCA1* and/or *BRCA2* (*BRCA1*/*BRCA2*) also confer selective sensitivity to PARP inhibitors. Thus, assays to detect *BRCA1*/*BRCA2*-deficient tumors have been sought. Recently, somatic substitution, insertion/deletion and rearrangement patterns, or 'mutational signatures', were associated with *BRCA1*/*BRCA2* dysfunction. Herein we used a lasso logistic regression model to identify six distinguishing mutational signatures predictive of *BRCA1*/*BRCA2* deficiency. A weighted model called HRDetect was developed to accurately detect *BRCA1*/*BRCA2*-deficient samples. HRDetect identifies *BRCA1*/*BRCA2*-deficient tumors with 98.7% sensitivity (area under the curve (AUC) = 0.98). Application of this model in a cohort of 560 individuals with breast cancer, of whom 22 were known to carry a germline *BRCA1* or *BRCA2* mutation, allowed us to identify an additional 22 tumors with somatic loss of *BRCA1* or *BRCA2* and 47 tumors with functional *BRCA1*/*BRCA2* deficiency where no mutation was detected. We validated HRDetect on independent cohorts of breast, ovarian and pancreatic cancers and demonstrated its efficacy in alternative sequencing strategies. Integrating all of the classes of mutational signatures thus reveals a larger proportion of individuals with breast cancer harboring *BRCA1*/*BRCA2* deficiency (up to 22%) than hitherto appreciated (~1–5%) who could have selective therapeutic sensitivity to PARP inhibition.**

A small fraction of breast cancers (~1–5%)[1–3] are attributed to familial mutations in the *BRCA1* and *BRCA2* cancer susceptibility genes. Heterozygous germline mutations in *BRCA1* and *BRCA2* confer elevated lifetime risks of breast, ovarian and other cancers[4,5]. BRCA1 and BRCA2 proteins have multiple, distinct roles in maintaining genome integrity, particularly through homologous recombination (HR)-mediated double-strand break (DSB) repair[6]. These classical tumor-suppressor genes usually lose the wild-type allele during tumorigenesis to become fully inactivated[7]. *BRCA1*- and *BRCA2*-null tumors are thus deficient in HR and are selectively sensitive to compounds that increase the demand on HR[8]. PARP inhibitors are an example of therapeutic compounds that cause replication fork stalling and collapse, leading to increased DSBs[9]. The inability to perform HR-dependent DSB repair ultimately leads to selective tumor cell death[10,11].

Preclinical studies and phase I and II breast and ovarian cancer clinical trials[12,13] have shown PARP inhibitor efficacy in familial *BRCA1*- and *BRCA2*-mutant patients. However, PARP inhibition has applications beyond the treatment of germline-mutated tumors[14]. Effective PARP inhibition maintenance therapy has been demonstrated in high-grade serous ovarian cancer with germline or somatic *BRCA1* or *BRCA2* mutations[15]. Thus, extensive efforts have been put into identifying the molecular features of tumors that are *BRCA1* or *BRCA2* deficient—a defect historically referred to as 'BRCAness'—whether the genes are inactivated through germline, somatic or secondary means, including promoter DNA hypermethylation or inactivation of a related gene in the HR pathway.

Gene-specific sequencing strategies, including sequencing all known HR genes, multiplex-ligation-dependent probe amplification (MLPA)[16], promoter hypermethylation assays[17], identification of transcriptional metagene signatures[18–20], copy-number-based methods (for example, to determine the homologous recombination deficiency (HRD) index and genomic 'scars')[21–23] and functional assays of HR competence[24], have been developed to detect *BRCA1*/*BRCA2* deficiency. However, the indices from these methods have had limited predictive success. A recent review suggests that a good predictor of the biological status of an HR-deficient tumor is essential, as the cohort of tumors that demonstrate BRCAness and could be selectively sensitive to PARP inhibitors is likely not limited to the small proportion of familial breast and ovarian cancers with *BRCA1* or *BRCA2* mutations but extends to a larger fraction of sporadic breast and ovarian cancers, as well as other cancer types[25].

Recent advances in sequencing technology[26] have greatly reduced sequencing costs, permitting whole-genome sequencing (WGS) for

the detection of all somatic mutations, including base substitutions, insertions/deletions (indels), rearrangements and copy number aberrations, in human cancers. Deep analysis reveals patterns of mutations, or somatic mutational signatures, which are the physiological readout of the DNA damage and DNA repair processes that have occurred through tumorigenesis[27–31]. These patterns are indicators of past and ongoing exposures, whether to environmental insults, such as UV radiation, or to endogenous biochemical degradation and deficiencies of DNA repair pathways like HR.

We reason that mutational signatures that report *BRCA1*/*BRCA2* deficiency in germline-mutated tumors could be used as a predictor of other tumors that also have this deficiency. Previously, base-substitution signature 3 was shown to distinguish germline *BRCA1*/*BRCA2*-null cancers from sporadic cancers in a small subset of breast cancers[29,30] and was subsequently extended to pancreatic[32,33], ovarian[34] and stomach[35] cancers. However, selecting a cutoff to discriminate *BRCA1*/*BRCA2*-deficient from *BRCA1*/*BRCA2*-proficient cancers is not straightforward when using this signature alone. Recent characterization of a large cohort of WGS breast cancers[27,28] has provided new insights into how these cancers can be distinguished. A defect in a single gene such as *BRCA1* or *BRCA2* does not result in a single signature—it gives rise to at least five mutational signatures of all classes, including base substitutions, indels and rearrangements[27,28]. Unlike most biomarkers, these multiple mutational signatures are the direct consequence of abrogation of DSB repair pathways. Thus, in the current analysis, we exploit this observation to quantitatively define the genomic features of *BRCA1*/*BRCA2* deficiency and present a WGS-based predictor with remarkable performance for detection of HR-deficient tumors.

## RESULTS

### Quantitatively defining features of BRCAness

Twenty-four individuals carrying inherited predisposition mutations in *BRCA1* ($n$ = 5) and *BRCA2* ($n$ = 19) were recruited into a breast cancer genome sequencing study involving 560 individuals[27]. Loss of the wild-type allele, which was predicted to result in complete inactivation of the relevant protein, was observed in 22 of the 24 breast cancers.

These 22 tumors had a distinguishing genomic profile: overrepresentation of base-substitution signature 3 or 8, an excess of large deletions (>3 bp) with microhomology at the junction of the deletion, rearrangement signature 5 and copy number profiles associated with the widespread loss of heterozygosity (**Fig. 1**). Additionally, *BRCA1*-null tumors mainly had an excess of rearrangement signature 3 mutations (characterized by short <10-kb tandem duplications) and a lesser contribution of rearrangement signature 1 mutations (typified by long >100-kb tandem duplications)[27].

The 22 *BRCA1*- or *BRCA2*-null tumors were used in a first training set to quantitatively define features of *BRCA1*/*BRCA2* deficiency. They were contrasted with a cohort of 235 cases of sporadic breast cancer with quiescent genomic profiles, which are distinct from the profiles of *BRCA1*/*BRCA2*-null cancers.

Somatic variants of all classes of mutation had been previously called. Twelve base-substitution, two indel and six rearrangement mutational signatures were previously extracted, and HRD copy number indices were obtained (**Supplementary Table 1**). A lasso logistic regression model was applied to counts of mutational signatures and HRD indices that were log transformed and normalized to permit comparability between genomic parameters (**Supplementary Table 2**).

An iterative tenfold nested cross-validation strategy was adopted, where 90% of samples were used for model parameter selection and the weights for each parameter were tested on the remaining 10% of samples. This was performed to ensure that the parameters identified as putative predictors of *BRCA1*/*BRCA2* deficiency were robust and generalizable.

Five distinguishing parameters with different individual weights were found to convey the greatest difference between *BRCA1*/*BRCA2*-deficient cancers and sporadic breast cancers: microhomology-mediated indels, the HRD index, base-substitution signature 3, rearrangement signature 3 and rearrangement signature 5 (**Supplementary Table 2**).

### Identification of additional *BRCA1*- and *BRCA2*-null tumors

The selected parameters were applied across the cohort of 560 breast cancers to test the performance of our model in predicting *BRCA1*/*BRCA2* deficiency and to detect other cancers with characteristics similar to germline *BRCA1*/*BRCA2*-null tumors (see **Fig. 2** for the workflow; **Supplementary Table 2** and **Supplementary Fig. 1**). The resulting distribution of probabilities of *BRCA1*/*BRCA2* deficiency was a strikingly steep sigmoidal curve with a clear distinction between the individuals predicted to have high and low probabilities of *BRCA1*/*BRCA2* deficiency. Apart from the 22 positive controls from the training set, 90 of 538 additional tumor samples were identified as having a probability of *BRCA1*/*BRCA2* deficiency exceeding 0.7, bringing the total proportion of patients predicted to have a high level of *BRCA1*/*BRCA2* deficiency to 20%.

This result prompted us to look for additional *BRCA1* and *BRCA2* mutations (germline and somatic) in the cohort of 560 patients. Thirty-three patients were found to carry pathogenic germline variants in *BRCA1* or *BRCA2* with corresponding somatic inactivation of the second allele. This more than doubles the number of individuals harboring familial cancer predisposition alleles relative to the number who were known to have germline *BRCA1* or *BRCA2* mutations when originally recruited into the study, carrying important clinical genetic counseling implications and potential for active surveillance and/or treatment choices for affected individuals and their families.

Twenty-two individuals had early, clonal somatic *BRCA1* or *BRCA2* mutations ($n$ = 8) or promoter DNA hypermethylation of *BRCA1* ($n$ = 14) with inactivation of the second allele. The remaining tumors with a probability of *BRCA1*/*BRCA2* deficiency exceeding 0.7 did not demonstrate biallelic inactivation of *BRCA1* or *BRCA2*, although DNA methylation data were not available for a subset of individuals.

Six *BRCA1*-null samples had probabilities of 0.006–0.64 and were missed because the algorithm had been trained on a small cohort of 5 *BRCA1*-mutant tumors of the total 22 in the training set, suggesting that algorithm retraining on a larger and more balanced cohort was prudent.

### HRDetect: predictor of *BRCA1*/*BRCA2* deficiency in cancer

Given that additional individuals were identified as null for *BRCA1* or *BRCA2*, we performed another iteration of the lasso logistic regression model on a larger, better-powered training set comprising 77 samples (22 with known germline mutations, 33 with new germline mutations and 22 with somatic mutations) (**Fig. 2**).

Reassuringly, the same genomic features were identified as predictive parameters that had been observed for the 22 germline-null samples, with the addition of base-substitution signature 8. These features, ranked by decreasing weight, included microhomology-mediated deletions (2.398), base-substitution signature 3 (1.611), rearrangement signature

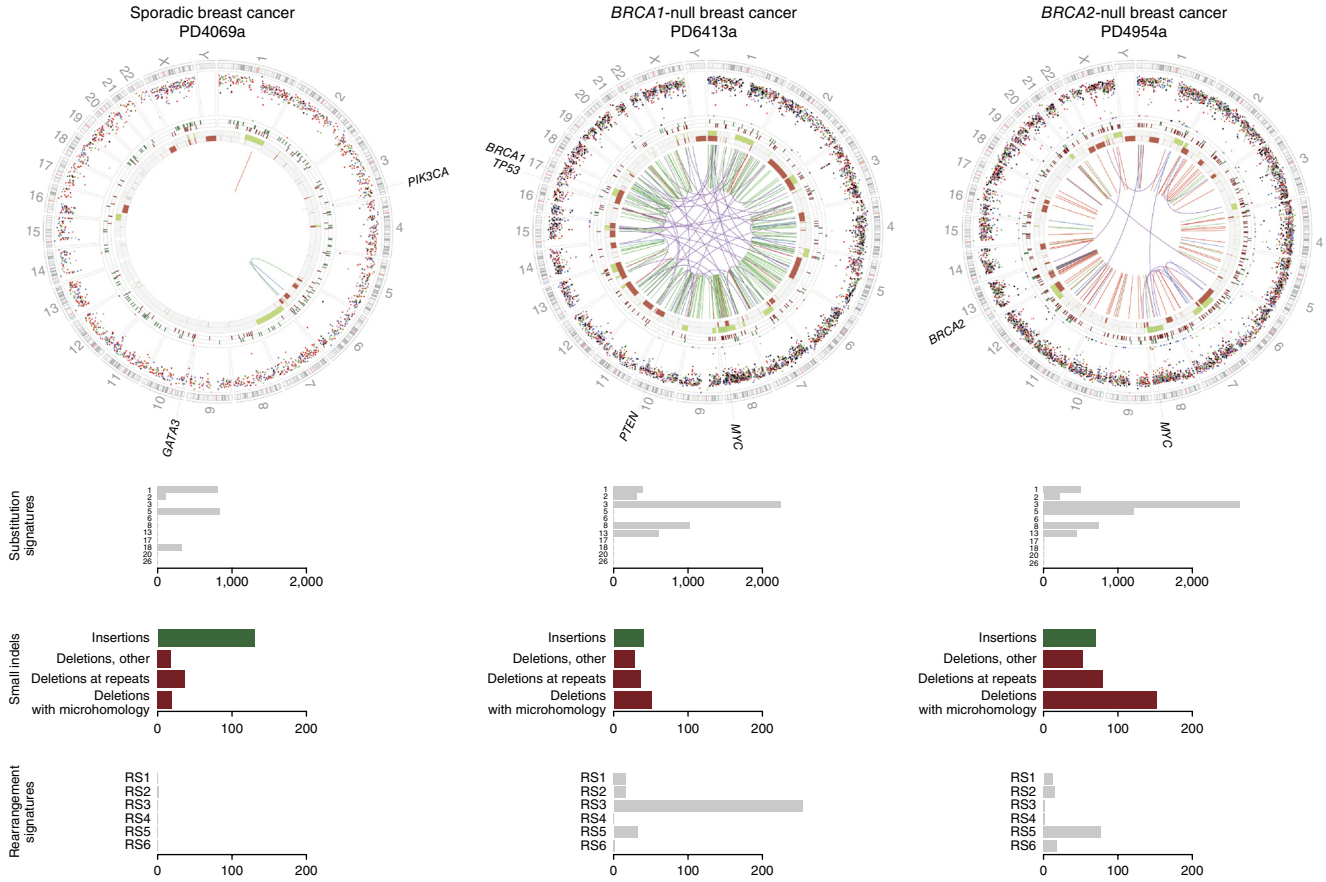



**Figure 1** Whole-genome profiling depicts differences between patients with *BRCA1*/*BRCA2*-mutated tumors and sporadic tumors. Examples of genome plots are shown for a typical sporadic breast cancer tumor (left), a *BRCA1* germline-null tumor (middle) and a *BRCA2* germline-null tumor (right). The features depicted in the Circos plots from the outermost rings moving inwards are (i) the karyotypic ideogram; (ii) base substitutions, plotted as rainfall plots ($\log_{10}$ (intermutation distance) on the radial axis; dot color: blue, C>A; black, C>G; red, C>T; gray, T>A; green, T>C; pink, T>G); (iii) insertions shown as short green lines; (iv) deletions shown as short red lines; (v) major (green blocks, gain) and minor (red blocks, loss) copy number alleles; and (vi) rearrangements shown as central lines (green, tandem duplications; red, deletions; blue, inversions; purple, interchromosomal events). Driver genes with mutations in breast cancer are labeled around each Circos plot. Below each Circos plot are histograms showing mutation counts for each mutation class: the topmost histogram shows the number of mutations contributing to each substitution signature; the middle histogram represents indel patterns; and the bottom histogram shows the number of rearrangements contributing to each rearrangement signature.

3 (1.153), rearrangement signature 5 (0.847), HRD index (0.667) and base-substitution signature 8 (0.091) (**Fig. 2b** and **Supplementary Table 3**). Acknowledging the imbalance in the numbers of *BRCA1*- or *BRCA2*-null cancers versus controls, supervised learning was repeated with a 1:1 ratio of these cases and controls. Differences between this assessment and the original one (77 null cancers:234 controls) were insignificant, verifying the stability of the six parameters critical for distinguishing *BRCA1*/*BRCA2*-deficient tumors. With a larger data set available, we permitted identification of interactions between genomic covariates in order to discover potentially augmented effects from cooperating signatures. Although correlations were observed (**Supplementary Fig. 2**), the performance including interactions did not improve on predictions made by the model without interactions. Therefore, we opted for the simpler model, keeping each genomic parameter independent. Thus, we finalized our predictor of *BRCA1*/*BRCA2* deficiency, termed HRDetect, on this set.

HRDetect was reapplied on the cohort of 560 cases of breast cancer (**Fig. 3**) and showed excellent performance, as demonstrated by a receiver operating characteristic (ROC) curve with an area under the curve (AUC) of 0.98 (**Fig. 4**). This result is unlikely to be bettered, emphasizing the value of utilizing multiple mutational signatures as a readout of *BRCA1*/*BRCA2* deficiency. Reinforcing this point, no individual genomic parameter performed as well as all six genomic signatures incorporated together in HRDetect (**Fig. 4**). In particular, HRDetect is superior to current methods of assessing *BRCA1*/*BRCA2* deficiency, specifically, genomic-scar-based indexes[21–23] such as the HRD score (sensitivity of HRD score = 0.6; the ROC curves in **Fig. 4** compare HRDetect against HRD score alone and other mutational signatures individually).

Using a probabilistic cutoff of 0.7, HRDetect predicted *BRCA1*/*BRCA2* deficiency with a sensitivity of 98.7% in the cohort of 560 individuals. HRDetect identified a total of 124 samples with a score exceeding 0.7, including an additional 47 samples with a high probability of *BRCA1*/*BRCA2* deficiency. These remaining tumors (5/340 estrogen receptor (ER)-positive and 42/143 ER-negative tumors) with high scores and neither germline nor somatic *BRCA1* or *BRCA2* mutations, and in which promoter hypermethylation of *BRCA1* was either not observed ($n$ = 10) or not available for assessment ($n$ = 37), were investigated for the inactivation of other genes involved in HR repair and for other germline susceptibility alleles.

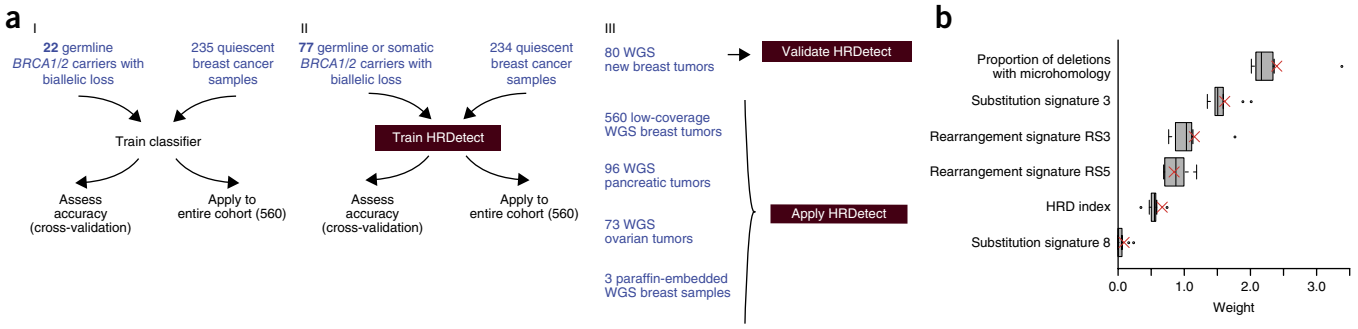



**Figure 2** Workflow for developing HRDetect. (**a**) Workflow of the steps involved in the development of the HRDetect predictor. (I) Initial training using 22 known germline *BRCA1*/*BRCA2*-null samples. (II) Retraining using 77 *BRCA1*/*BRCA2*-null samples to produce the final HRDetect predictor. (III) Validation on a further set of breast cancers and application to other data sets. (**b**) Box plots of the weights for the genomic features contributing to the HRDetect predictor. The range of values from 10 replicates of training in cross-validation using 311 breast cancer samples (77 *BRCA1*/*BRCA2*-null samples and 234 quiescent tumors) is shown. Red crosses indicate the final weights used in HRDetect. In the box plots, the midline represents the median, the two edges of the box represent the lower and upper interquartile range (IQR), upper whisker = min(max($x$), $Q_3 + 1.5 \times IQR$) and lower whisker = max(min($x$), $Q_1 - 1.5 \times IQR$), and the dots are outliers beyond the whiskers.

## Other genetic factors and BRCAness

Of these 47 samples with high HRDetect scores and without biallelic *BRCA1* or *BRCA2* mutations, 3 had mutations in HR genes. One individual, PD4875a, exhibited a high HRDetect score (0.94) and had a profile that is typically associated with *BRCA2* nullness. Although she carried a germline *BRCA2* mutation, the other parental allele was retained[36]. This patient was thus the exception where genetic *BRCA2* nullness could not be proven in the tumor. Inactivation of the wild-type allele by alternative means cannot be excluded. This patient also carried a germline truncating mutation in *BRIP1* (a gene associated with moderate penetrance breast cancer risk) with loss of the second allele; however, with only a single example of a truncating *BRIP1* mutation, the significance of this mutation is unclear. Otherwise, all other tumors with monoallelic germline or somatic inactivation of *BRCA1* or *BRCA2* were associated with low HRDetect scores (13 individuals in total: 4 with germline inactivation, 7 with somatic inactivation and 2 with *BRCA1* promoter hypermethylation) (**Supplementary Table 4**).

Two individuals (PD24205a and PD24212a) had somatic monoallelic truncating mutations in *PALB2* (also associated with moderate penetrance breast cancer risk). A third individual, PD11340a, who had a low HRDetect score also had a deleterious somatic monoalleleic essential-splice-site *PALB2* mutation. Given the small numbers of individuals with mutations in this gene, we would interpret the contribution of *PALB2* mutations with caution, as other modes of *BRCA1*/*BRCA2* inactivation or inactivation of other genes related to the HR pathway could underlie the observations in these patients.

Interestingly, monoallelic somatic inactivating mutations of other HR repair genes including *ATR* (PD14457a, PD23564a, PD5956a) and *ATM* (PD5937a) were not associated with high scores of *BRCA1*/*BRCA2* deficiency. Furthermore, none of the genes from the list of HR genes (*RAD51C*, *RAD50*, *CHEK2* and *FANCA*–*PALB2* (*FANCN*), the *FANC* group of genes) was identified as a contributor among tumors with high HRDetect scores. Notably, high- and moderate-penetrance germline breast cancer susceptibility alleles, including for *TP53*, *PTEN*, *ATM*, *CHEK2*, *ATR*, *RAD50*, *CDH1*, *STK11* and *PALB2*, whether in the context of monoallelic or biallelic inactivation, were not associated with either a genomic profile or a high probability of *BRCA1*/*BRCA2* deficiency.

These results first emphasize the importance of knowing the status of the alternative parental allele in the interpretation of mutation data (**Supplementary Table 4**). Second, they highlight that nearly one-third of tumors with high HRDetect scores predicting *BRCA1*/*BRCA2* deficiency cannot be authenticated as *BRCA1* or *BRCA2* null through genetic and/or epigenetic means. Yet, given the striking resemblance of these tumors to *BRCA1*- and *BRCA2*-null tumors, it is intriguing to consider that these cancers are biologically comparable and likely to respond similarly as *BRCA1*- and *BRCA2*-null cancers, particularly to PARP inhibition.

## Validation of HRDetect in a new cohort of 80 WGS breast cancers

As a validation exercise, HRDetect was applied to a new cohort of WGS breast cancers that were mainly ER positive and HER2 negative from 80 individuals (**Fig. 4** and **Supplementary Table 5**). HRDetect successfully identified one germline *BRCA1* mutation carrier and five *BRCA2* mutation carriers (four germline and one somatic) with associated loss of the wild-type allele. One individual, PD14434a, carried a germline essential-splice-site mutation with loss of the alternative parental allele but fell short of the HRDetect cutoff of 0.7. Two samples, one with a germline and one with a somatic *BRCA2* mutation, retained the alternative allele and were correctly assigned low HRDetect scores. Thus, the sensitivity of HRDetect on this validation cohort was high at 86%.

## Performance of HRDetect with alternative sequencing strategies

To explore the performance of HRDetect with alternative sequencing strategies, we performed an *in silico* experiment, randomly downsampling the sequences of the 560 high-coverage (30- to 40-fold) WGS breast cancers in order to generate low-coverage (10-fold; range 9.9- to 10.5-fold) WGS sequence files for analysis (**Supplementary Table 6**). Somatic mutations were called across the downsampled sequences, signatures and HRD indices were extracted and the performance of HRDetect was tested. In theory, the absolute detection of every somatic change is not obligatory, as long as some mutations that are representative of the overarching mutation patterns are present.

As expected, the numbers of base substitutions, indels and rearrangements were consistently lower in the downsampled *in silico* experiment of all samples when compared to the original high-coverage experiment. Nevertheless, all 12 base-substitution, 2 indel and 6 rearrangement signatures were detectable, at approximately the same proportions per sample, albeit they were present at reduced absolute numbers (**Supplementary Fig. 3**). Additionally, copy number

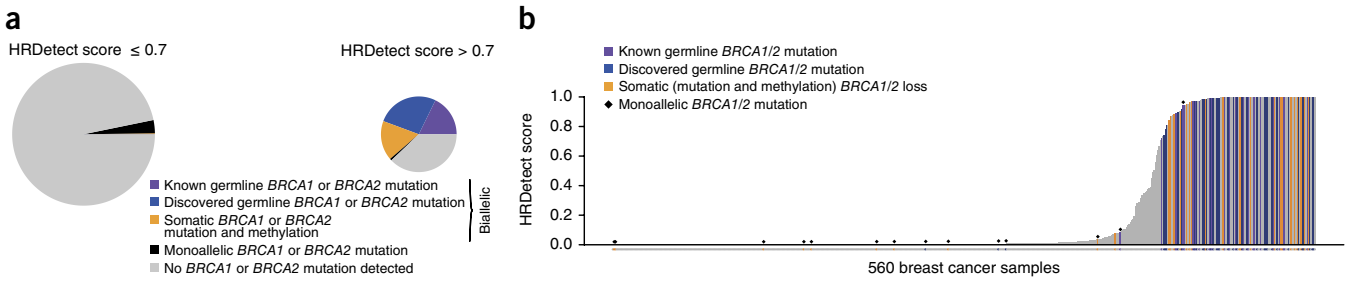



**Figure 3** HRDetect as a probabilistic classifier. (**a**) Pie charts depicting the *BRCA1* and *BRCA2* mutation status of samples in the data set of 560 breast cancers. Left, samples with HRDetect scores below the cutoff of 0.7. Right, samples with HRDetect scores above the cutoff of 0.7. Purple, germline *BRCA1*/*BRCA2* mutation carriers with loss of the second allele, previously known at recruitment into the study; blue, newly discovered germline *BRCA1*/*BRCA2* mutation carriers with loss of the second allele; orange, carriers of somatic germline *BRCA1*/*BRCA2* mutation and/or DNA hypermethylation of the *BRCA1* promoter with loss of the second allele; black, monoallelic germline or somatic *BRCA1*/*BRCA2* mutation carriers retaining the second allele; gray, samples in which no *BRCA1*/*BRCA2* mutation has been detected. (**b**) The HRDetect scores of 560 breast cancer samples ordered from lowest to highest score across the *x* axis from left to right. Colored bars represent both samples with monoalleleic mutations and those with loss of the second allele: purple, previously known carriers of germline *BRCA1*/*BRCA2* mutations (24 in total, of which 22 have biallelic loss and 2 have monoalleleic loss); blue, newly discovered carriers of germline *BRCA1*/*BRCA2* mutation (36 in total, of which 33 have biallelic loss and 3 have monoalleleic loss); orange, carriers of somatic germline *BRCA1*/*BRCA2* mutations and DNA hypermethylation of the *BRCA1* promoter (31 in total, of which 22 have biallelic loss and 9 have monoalleleic loss); black diamonds above the bars indicate samples with monoallelic germline or somatic *BRCA1*/*BRCA2* mutant alleles that retained the second allele (14 in total).

analysis showed good concordance for overall HRD scores ($r = 0.63$) between the high- and low-coverage genomes. Thus, despite the reduction in sensitivity for individual somatic mutations, the detection of overarching mutation signatures remained relatively secure.

At an absolute probability cutoff of 0.7, the sensitivity for detection of *BRCA1*/*BRCA2*-defective cancers in a low-coverage genome sequencing experiment remained high at 86% (**Fig. 4**). The concordance in HRDetect predictions between high- and low-coverage sequencing experiments was excellent ($r = 0.96$). Low-coverage genome sequences may thus be adequate for HRDetect to report deficiency of *BRCA1*/*BRCA2*.

In contrast, when HRDetect was used to assess *BRCA1*/*BRCA2* deficiency in data that were representative of only coding sequences (whole-exome sequencing, WES), the sensitivity of detection was affected considerably, falling to 46.8%. This is because data on essential predictor components such as rearrangement signatures 3 and 5 are not available by WES, and substitutions/indels are restricted to only 1–1.5% of the footprint of the genome (**Supplementary Note**). When the HRDetect algorithm was retrained taking WES-based data alone as input, the performance of the classifier was improved (sensitivity, 73%, correctly identifying 56 of 77 *BRCA1*- or *BRCA2*-null tumors), although at the cost of calling 12 additional samples that were not previously identified as *BRCA1*/*BRCA2* deficient (**Supplementary Table 7**).

## Application of HRDetect to predict BRCAness in other types of cancer

HRDetect was applied to other WGS cancers, including pancreatic and ovarian cancers, to assess the predictor's generalizability across tumor types[32–34]. Available BAM files were analyzed using our somatic mutation calling pipeline, mutational signatures were extracted and copy number profiles were obtained.

The ovarian cancer cohort comprised 73 samples. Using a threshold of 0.7, 46 (63%) were identified as having a high probability of *BRCA1*/*BRCA2* deficiency. Of these, 30 were confirmed as having germline or somatic *BRCA1* or *BRCA2* mutations with loss of the wild-type parental allele (14 with germline mutations, 6 with somatic mutations and 10 with altered DNA methylation) (**Supplementary Table 8**). No *BRCA1*/*BRCA2*-null ovarian samples were found to have a low HRDetect score. Thus, again, HRDetect has a sensitivity of detecting *BRCA1*/*BRCA2*-null cancers approaching 100%, and it uncovered 16 additional individuals in the ovarian cancer cohort as HR deficient.

The pancreatic cancer cohort comprised 96 samples. Eleven (11.5%) were found to have a high HRDetect score. Five were mutated for *BRCA1* or *BRCA2* (three with germline mutations and two with somatic mutations) and had lost the wild-type allele, and one had retained the wild-type allele. Epigenetic data were not available to interrogate the status of the remaining five samples. Three samples had *BRCA2* mutations, but these did not show convincing evidence of loss of the wild-type allele and had low HRDetect scores (**Supplementary Table 8**). Thus, HRDetect had a sensitivity approaching 100% in this pancreatic cancer cohort and identified five additional patients with potential HR deficiency.

Overall, HRDetect had excellent sensitivity for these other tumor types. However, the distributions of HRDetect scores were slightly different (**Fig. 4c**). When more samples become available, thus increasing power for analysis, a reappraisal of HRDetect parameters per tissue type may be necessary to fine-tune performance in different tissue types.

## Strengths of HRDetect scores and their relevance to accelerating clinical application

Routine clinical pathology practice involves the storage of tumor material using formalin fixation and paraffin embedding (FFPE) methods. To explore the performance of HRDetect on FFPE tissue samples (**Fig. 5**), we obtained nucleic acids derived from an FFPE sample from a patient with a germline *BRCA1* mutation. WGS was performed, somatic mutations were called and mutational signatures were extracted. HRDetect correctly reported a high probability (0.94) of *BRCA1*/*BRCA2* deficiency, despite an overwhelming FFPE-related sequencing artifact (**Fig. 5** and **Supplementary Table 9**) that compromised substitution signature extraction, resulting in the absence of base-substitution signature 3. Small amounts of the correct combination of other critically distinguishing signatures still generated a strong probabilistic prediction.

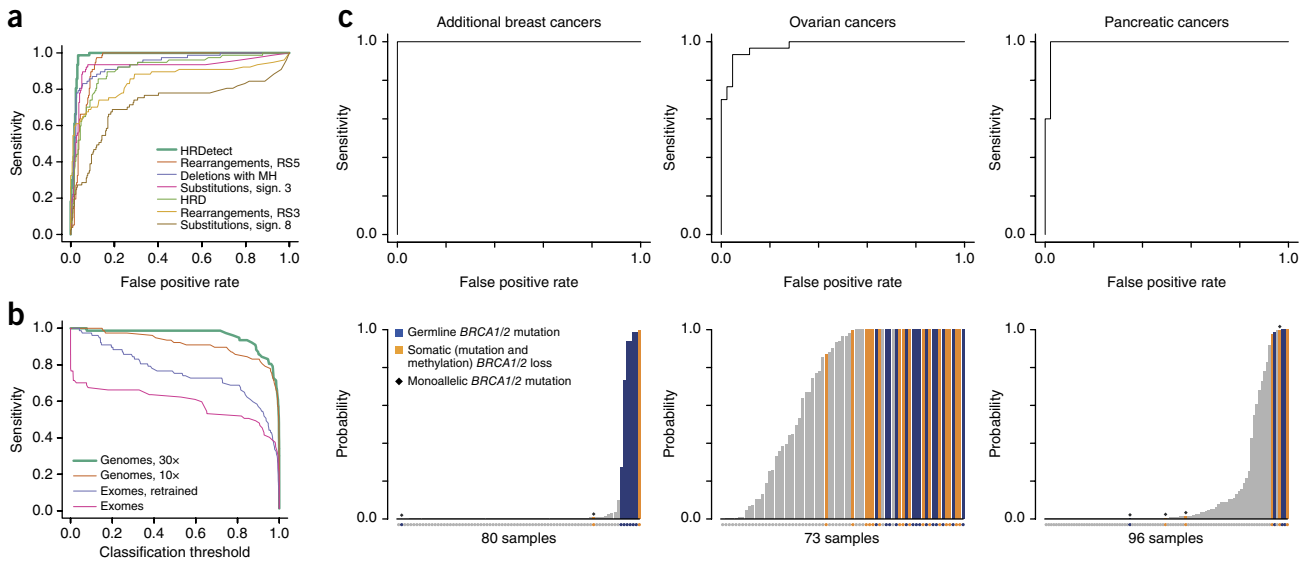



**Figure 4** Performance of HRDetect and validation. (**a**) ROC curves demonstrating the performance of HRDetect on 371 breast cancer samples, as well as its performance when using individual mutational signatures as a predictor of *BRCA1*/*BRCA2* deficiency. MH, microhomology. (**b**) Comparison of the sensitivity of detection of *BRCA1*/*BRCA2*-deficient tumors across different types of sequencing experiments: high-coverage (30–40×) genomes; low-coverage (10×) genomes; and WES (using HRDetect weights learned from WGS data and retrained on WES data). Samples from 371 breast cancers were used in each case. (**c**) Performance of HRDetect on other data sets. From left to right: cohorts of 80 new breast cancers, 73 ovarian cancers and 96 pancreatic cancers. Top, ROC curves for each cancer type. Bottom, histograms of HRDetect scores. Blue, carriers of germline *BRCA1*/*BRCA2* mutations; orange, carriers of somatic *BRCA1*/*BRCA2* mutations; gray, samples with no *BRCA1*/*BRCA2* mutation detected; black diamonds above the bars indicate samples with monoallelic germline or somatic *BRCA1*/*BRCA2*-mutant alleles that retained the second allele.

Biological hypermutation phenomena occur in human cancers, and mutational processes, such as those due to the APOBEC family of cytidine deaminases, are not uncommon. We found that *BRCA1*/*BRCA2*-deficient cancers remained consistently identified by HRDetect despite excessive APOBEC-related mutagenesis in some samples (see **Supplementary Fig. 4** for an example). Furthermore, HRDetect was able to discern *BRCA1*/*BRCA2*-deficient cancers with remarkable precision over a wide range of tumor cellularities, including for samples with relatively low cellularity (but not less than 15%) (**Supplementary Fig. 5**), where mutation calling sensitivity may be compromised. Thus, irrespective of biological or nonbiological noise, HRDetect faithfully detects the signal of *BRCA1*/*BRCA2* deficiency, reinforcing the exceptional utility of this classifier.

Finally, to advance the potential clinical utility of HRDetect, we considered whether HRDetect could be applied earlier in the clinical process on small needle biopsy samples rather than post-operatively on large specimens. To this end, we obtained 18 DNA samples (14 needle biopsies and 4 postoperative tumor block specimens) from 9 individuals with triple-negative tumors who were treated with neoadjuvant anthracyclines with or without taxanes[37] (**Supplementary Table 9** and **Supplementary Note**). Although anthracyclines are a different compound from PARP inhibitors, sensitivity to anthracyclines has been reported for tumors that show *BRCA1*/*BRCA2* deficiency[38,39]. Interestingly, four patients demonstrated complete responses to treatment, and all had high HRDetect scores—two were confirmed to be germline *BRCA1* mutation carriers, and two had sporadic tumors (**Fig. 5c**). In contrast, five patients who exhibited residual disease had low HRDetect probability scores. Furthermore, HRDetect performed consistently in independent biopsies of tumors and in comparison of biopsy and postoperative specimens taken from the same patient, without exception. Although the numbers of samples are small, in all, these analyses suggest that HRDetect has the potential to distinguish therapeutic sensitivity as early in the patient's clinical journey as the first biopsy and is robust between independent biopsies and/or specimens. Larger clinical trials are clearly necessary to fully understand how this predictor will perform when applied to breast cancer diagnostics in general.

## Variants of uncertain significance

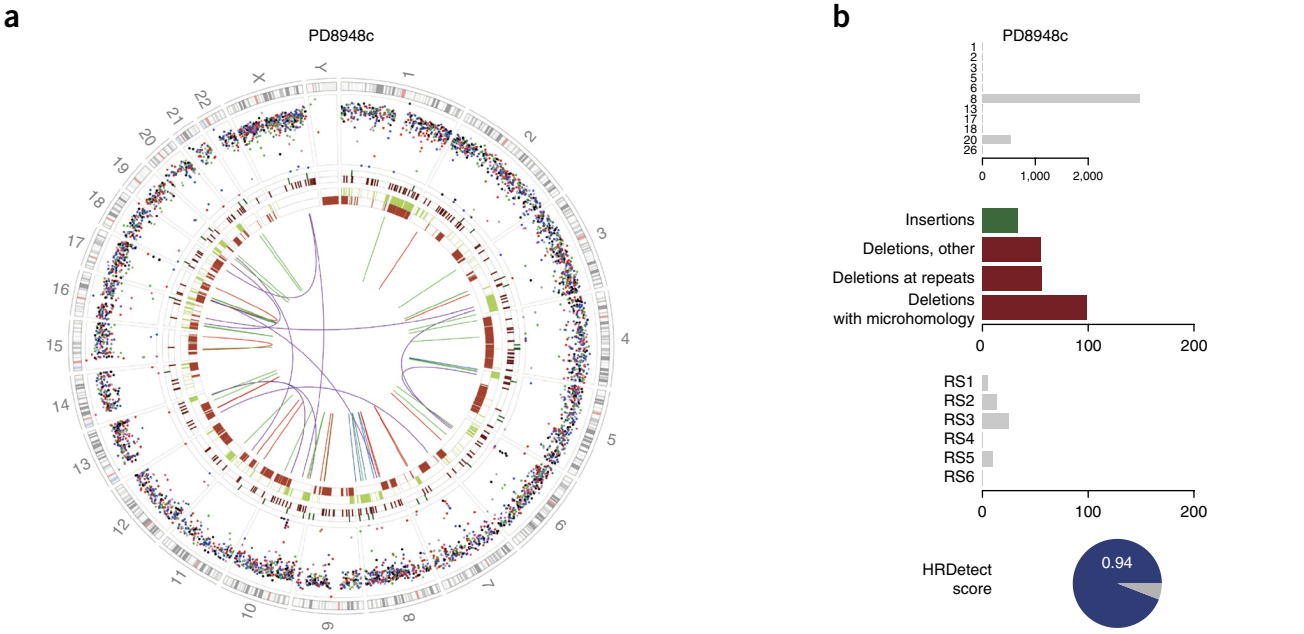
Germline *BRCA1* and *BRCA2* SNPs and variants of uncertain significance (VUSs), including 20 common alleles and 107 rare or private variants, were identified in the data set comprising 560 breast cancers (**Supplementary Table 4**). Fifty-six had concurrent loss of the second allele. However, tumors with these variants did not consistently demonstrate high scores for *BRCA1*/*BRCA2* deficiency, emphasizing that these VUSs are unlikely to be pathogenic and hence are of low clinical significance.

There was one exception: PD23563a had a missense mutation in *BRCA1* encoding p.L1780P with loss of heterozygosity of the other allele and a high score for *BRCA1*/*BRCA2* deficiency. This variant remains "of uncertain significance" in clinical databases of *BRCA1* SNP alleles (ClinVar), although functional support for defective BRCA1 function has been reported[40]. With only a single example, this result must be interpreted with caution, as other causes of *BRCA1*/*BRCA2* deficiency in this sample cannot be excluded.

Additionally, eight missense somatic *BRCA1* and *BRCA2* mutations were identified and did not appear to be associated with features of deficiency. Over time, HR mutational signatures could be used to effectively validate the pathogenicity of VUSs.

## HRDetect can distinguish *BRCA1*- from *BRCA2*-deficient tumors

Thus far, we have focused on detecting tumors with either *BRCA1* or *BRCA2* deficiency and not distinguishing between them. Currently, there is no clinical indication to separately identify these tumors because *BRCA1*- and *BRCA2*-deficient tumors are similarly sensitive to PARP inhibition. In the future, however, reasons for separating




c

| Patient | PD9768 | PD9769 | PD9770 | PD9771 | PD9777 | PD9772 | PD9774 | PD9775 | PD9776 |
|---|---|---|---|---|---|---|---|---|---|
| treatment response | RD | RD | RD | RD | RD | CR | CR | CR | CR |
| *BRCA1*/*BRCA2* status | | | | | | | BRCA1 | | BRCA1 |
| Pretreatment biopsy 1 | PD9768c 0.02 | PD9769c 0.22 | PD9770c 0.01 | PD9771c 0.00 | PD9777c 0.55 | PD9772a2 1.00 | PD9774c 1.00 | PD9775c 1.00 | PD9776c 1.00 |
| Pretreatment biopsy 2 | | | PD9770d 0.14 | PD9771a 0.01 | PD9777a 0.03 | | | PD9775a 1.00 | PD9776a 1.00 |
| Post-treatment specimen from surgical block | | PD9769a 0.16 | PD9770a 0.39 | PD9771d 0.03 | PD9777d 0.07 | | | | |

**Figure 5** Clinically relevant strengths of HRDetect. (**a**) Genome plot of an FFPE sample from a patient with a germline *BRCA1* mutation. Plot elements are as described in **Figure 1**. (**b**) Contribution of mutation signatures. Top, substitutions; middle, indels; bottom, rearrangements. A representation of the HRDetect score is shown below. (**c**) HRDetect scores for nine patients treated with neoadjuvant anthracyclines with or without taxanes. Duplicate pretreatment needle biopsy samples were available for five of the patients (pretreatment biopsies 1 and 2). One patient (PD9770) had multifocal tumors. One patient with extremely low tumor cellularity in both biopsies and with hardly any mutations was excluded (PD9773). HRDetect scores are provided under the patient ID for each sample. Blue shading indicates patients with residual disease (RD) after treatment, while patients shaded green had a complete response to treatment (CR).

these tumors may arise. The discriminating genomic parameters that distinguish *BRCA1*- from *BRCA2*-deficient tumors are rearrangement signature 3 and deletions without distinctive junctional characteristics (**Supplementary Note** and **Supplementary Fig. 6**). Both are best detected using WGS approaches.

## DISCUSSION

Abrogation of *BRCA1* and *BRCA2* leads to not only one mutational signature but a characteristic set of these signatures. As a predictive tool, utilizing these multiple pathogenomic mutational signatures is extraordinarily effective, with performance metrics suggesting that this method cannot be easily bettered. It dependably detects *BRCA1*/*BRCA2* deficiency in the presence of biological noise (for example, APOBEC-related mutagenesis), when there is relatively low tumor cellularity (**Supplementary Note**) and when there is nonbiological noise from formalin fixation. Demonstrating HRDetect efficacy in FFPE-banked samples opens doors in terms of the exploration of historic and/or existing clinical trials, assuming that matched normal DNA samples are also available. Our analyses also emphasize that a WGS approach (even at low coverage (tenfold)) is far more effective than a WES approach at detecting *BRCA1*/*BRCA2* deficiency. Additionally, distinguishing *BRCA1*- from *BRCA2*-deficient tumors is dependent on WGS-based methods. These methodological points have implications for the design of the genomic aspects of clinical trials. At least for detecting *BRCA1*/*BRCA2* deficiency, WGS approaches are optimal. Our analyses provide support and context for large-scale national WGS endeavors such as the UK 100,000 Genomes Project and the Precision Medicine Initiative in the United States (see URLs).

While the performance of HRDetect is extremely promising, algorithmic developments are also envisaged, including for the identification

of tumors that have developed treatment-resistant alleles[41,42]. Historic scars of HR deficiency will be present in a resistant tumor, possibly leading to high HRDetect scores. However, distinguishing ongoing from historic mutational signatures of *BRCA1*/*BRCA2* deficiency is already a possibility, given the advances in exploiting the digital nature of modern sequencing technologies to construct phylogenetic trees of each person's tumor[30,37].

Although only 22 patients were originally recruited with known germline *BRCA1*- or *BRCA2*-null cancers, HRDetect reveals an additional 33 tumors with a germline mutation, 22 tumors with a somatic mutation and 47 tumors where no mutation was detected—bringing the total number of *BRCA1*/*BRCA2*-deficient tumors to 124 (22%). Large-scale population-based studies are required to gather proper population estimates, but, nevertheless, the numbers are startling.

Most notably, knowledge of the precise causative mutation may not be necessary because mutational signatures are such a reliable reporter of a tumor's biological status and, hence, its possible sensitivity to PARP inhibition (or other treatments—for example, platinum-based salts, anthracyclines and mitomycin C—to which cancers with BRCAness are selectively sensitive). Nearly one-third of samples that have characteristic genome profiles and high scores for *BRCA1*/*BRCA2* deficiency do not have canonical mutations detected. Thus, limiting testing to simply sequencing the *BRCA1* and *BRCA2* genes and performing methylation assays would miss this cohort of patients that could have functional deficiency incurred through currently unknown means.

If the tumors with predicted *BRCA1*/*BRCA2* deficiency also demonstrate sensitivity to PARP inhibitors, this would unearth a substantial cohort of patients who could be responsive to selective therapeutic agents, which are currently reserved for just ~1–5% of patients with breast cancer who are germline mutation carriers. This is potentially transformative, and thus the application of this predictor in PARP inhibitor clinical trials is warranted to assess predictive capacity in clinical settings.

The primary investment of a bank of WGS cancer data has been vital to the development of this predictor. Being able to find definitive ways of classifying the biological status of a patient's tumor, potentially for therapeutic stratification, is an example of the added value derived from these data—instrumental early steps that ultimately could lead to population health economic benefits.

**URLs.** ClinVar, http://www.ncbi.nlm.nih.gov/clinvar/; 100,000 Genomes Project, https://www.genomicsengland.co.uk/the-100000-genomes-project/; Precision Medicine Initiative, http://www.cancer.gov/research/key-initiatives/precision-medicine.

## METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*


### ACKNOWLEDGMENTS

This work has been performed on data that were previously published. They were generated and funded through the ICGC Breast Cancer Working group by the Breast Cancer Somatic Genetics Study (BASIS), a European research project funded by the European Community's Seventh Framework Programme (FP7/2010-2014) under grant agreement number 242006; the Triple Negative project funded by the Wellcome Trust (grant reference 077012/Z/05/Z) and the HER2+ project funded by Institut National du Cancer (INCa) in France (grants 226-2009, 02-2011, 41-2012, 144-2008, 06-2012). The ICGC Asian Breast Cancer Project was funded through a grant of the Korean Health Technology R&D Project, Ministry of Health & Welfare, Republic of Korea (A111218-SC01). The Oslo Breast Cancer Research Consortium (OSBREAC), Norway (http://www.osbreac.no/), contributed samples to the study. D.G. was supported by the EU-FP7-SUPPRESSTEM project. A.L.R. is partially supported by the Dana-Farber/Harvard Cancer Center SPORE in Breast Cancer (NIH/NCI 5 P50 CA168504-02). A.S. was supported by Cancer Genomics Netherlands (CGC.nl) through a grant from the Netherlands Organisation of Scientific research (NWO). C.S. is supported by a grant from the Breast Cancer Research Foundation. E.B. was funded by EMBL. A.T. acknowledges infrastructure support funding from the NIHR Biomedical Research Centres at Guy's and St Thomas' and Royal Marsden Hospital NHS Foundation Trusts. G.K. is supported by National Research Foundation of Korea (NRF) grants funded by the Korean government (NRF 2015R1A2A1A10052578). S.N.-Z. is a Wellcome Beit Fellow and personally funded by a Wellcome Trust Intermediate Fellowship (WT100183MA). Finally, we would like to acknowledge all members of the ICGC Breast Cancer Working Group and ICGC Asian Breast Cancer Project, for without the foresight of engaging in this scale of collaboration we would not have gained these insights.


### AUTHOR CONTRIBUTIONS

H.D., D.G. and S.N.-Z. drove the development of the intellectual concepts, performed analyses and wrote the manuscript. S. Morganella, J.S., X.Z. and M.R. contributed towards data curation and performed analyses. L.R.Y., S.B., A.M.S., P.T.S., T.A.K., J.E.E., P.N.S., S.R.L., A.V.-S., C.S., A.T., A.M.T. and S.V.L. contributed new samples and/or to experimental design of the study. S. Martin was the scientific project coordinator. K.R. provided bioinformatics support. P.J.C. provided infrastructure at the Wellcome Trust Sanger Institute. G.K., A.B., E.B., H.G.S., M.J.v.d.V., A.-L.B.-D., J.W.M.M., A.M.T., A.L.R., A.V. and M.R.S. originally conceived the concept of the Breast Cancer Consortium that generated the data resource that has been utilized for these analyses, contributed old and new samples, and contributed comments towards the manuscript.

### COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.


1. Anglian Breast Cancer Study Group. Prevalence and penetrance of *BRCA1* and *BRCA2* mutations in a population-based series of breast cancer cases. *Br. J. Cancer* **83**, 1301–1308 (2000).
2. John, E.M. *et al.* Prevalence of pathogenic *BRCA1* mutation carriers in 5 US racial/ethnic groups. *J. Am. Med. Assoc.* **298**, 2869–2876 (2007).
3. Malone, K.E. *et al.* Prevalence and predictors of *BRCA1* and *BRCA2* mutations in a population-based study of breast cancer in white and black American women ages 35 to 64 years. *Cancer Res.* **66**, 8297–8308 (2006).
4. Couch, F.J., Nathanson, K.L. & Offit, K. Two decades after BRCA: setting paradigms in personalized cancer care and prevention. *Science* **343**, 1466–1470 (2014).
5. King, M.C. "The race" to clone *BRCA1*. *Science* **343**, 1462–1465 (2014).
6. Lord, C.J. & Ashworth, A. The DNA damage response and cancer therapy. *Nature* **481**, 287–294 (2012).
7. Venkitaraman, A.R. Cancer suppression by the chromosome custodians, BRCA1 and BRCA2. *Science* **343**, 1470–1475 (2014).
8. Farmer, H. *et al.* Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434**, 917–921 (2005).
9. Prakash, R., Zhang, Y., Feng, W. & Jasin, M. Homologous recombination and human health: the roles of BRCA1, BRCA2, and associated proteins. *Cold Spring Harb. Perspect. Biol.* **7**, a016600 (2015).
10. Bryant, H.E. *et al.* Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature* **434**, 913–917 (2005).
11. Fong, P.C. *et al.* Inhibition of poly(ADP-ribose) polymerase in tumors from *BRCA* mutation carriers. *N. Engl. J. Med.* **361**, 123–134 (2009).
12. Audeh, M.W. *et al.* Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with *BRCA1* or *BRCA2* mutations and recurrent ovarian cancer: a proof-of-concept trial. *Lancet* **376**, 245–251 (2010).
13. Tutt, A. *et al.* Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with *BRCA1* or *BRCA2* mutations and advanced breast cancer: a proof-of-concept trial. *Lancet* **376**, 235–244 (2010).
14. Mateo, J. *et al.* DNA-repair defects and olaparib in metastatic prostate cancer. *N. Engl. J. Med.* **373**, 1697–1708 (2015).
15. Ledermann, J. *et al.* Olaparib maintenance therapy in platinum-sensitive relapsed ovarian cancer. *N. Engl. J. Med.* **366**, 1382–1392 (2012).
16. Lips, E.H. *et al.* Quantitative copy number analysis by Multiplex Ligation–dependent Probe Amplification (MLPA) of *BRCA1*-associated breast cancer regions identifies BRCAness. *Breast Cancer Res.* **13**, R107 (2011).

17. Ruscito, I. *et al. BRCA1* gene promoter methylation status in high-grade serous ovarian cancer patients—a study of the Tumour Bank Ovarian Cancer (TOC) and Ovarian Cancer Diagnosis consortium (OVCAD). *Eur. J. Cancer* **50**, 2090–2098 (2014).
18. Jazaeri, A.A. *et al.* Gene expression profiles of *BRCA1*-linked, *BRCA2*-linked, and sporadic ovarian cancers. *J. Natl. Cancer Inst.* **94**, 990–1000 (2002).
19. Larsen, M.J. *et al.* Classifications within molecular subtypes enables identification of *BRCA1*/*BRCA2* mutation carriers by RNA tumor profiling. *PLoS One* **8**, e64268 (2013).
20. Peng, G. *et al.* Genome-wide transcriptome profiling of homologous recombination DNA repair. *Nat. Commun.* **5**, 3361 (2014).
21. Joosse, S.A. *et al.* Prediction of *BRCA1*-association in hereditary non-*BRCA1*/*2* breast carcinomas with array-CGH. *Breast Cancer Res. Treat.* **116**, 479–489 (2009).
22. Vollebergh, M.A. *et al.* An aCGH classifier derived from *BRCA1*-mutated breast cancer and benefit of high-dose platinum-based chemotherapy in HER2-negative breast cancer patients. *Ann. Oncol.* **22**, 1561–1570 (2011).
23. Watkins, J.A., Irshad, S., Grigoriadis, A. & Tutt, A.N. Genomic scars as biomarkers of homologous recombination deficiency and drug response in breast and ovarian cancers. *Breast Cancer Res.* **16**, 211 (2014).
24. Graeser, M. *et al.* A marker of homologous recombination predicts pathologic complete response to neoadjuvant chemotherapy in primary breast cancer. *Clin. Cancer Res.* **16**, 6159–6168 (2010).
25. Lord, C.J. & Ashworth, A. BRCAness revisited. *Nat. Rev. Cancer* **16**, 110–120 (2016).
26. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
27. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
28. Morganella, S. *et al.* The topography of mutational processes in breast cancer genomes. *Nat. Commun.* **7**, 11383 (2016).
29. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
30. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
31. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
32. Waddell, N. *et al.* Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495–501 (2015).
33. Bailey, P. *et al.* Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **531**, 47–52 (2016).
34. Patch, A.M. *et al.* Whole-genome characterization of chemoresistant ovarian cancer. *Nature* **521**, 489–494 (2015).
35. Alexandrov, L.B., Nik-Zainal, S., Siu, H.C., Leung, S.Y. & Stratton, M.R. A mutational signature in gastric cancer suggests therapeutic strategies. *Nat. Commun.* **6**, 8683 (2015).
36. Stefansson, O.A. *et al.* Genomic and phenotypic analysis of *BRCA2* mutated breast cancers reveals co-occurring changes linked to progression. *Breast Cancer Res.* **13**, R95 (2011).
37. Yates, L.R. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21**, 751–759 (2015).
38. Rodriguez, A.A. *et al.* DNA repair signature is associated with anthracycline response in triple negative breast cancer patients. *Breast Cancer Res. Treat.* **123**, 189–196 (2010).
39. Chappuis, P.O. *et al.* A significant response to neoadjuvant chemotherapy in *BRCA1/2* related breast cancer. *J. Med. Genet.* **39**, 608–610 (2002).
40. Lee, M.S. *et al.* Comprehensive analysis of missense variations in the BRCT domain of BRCA1 by structural and functional assays. *Cancer Res.* **70**, 4880–4890 (2010).
41. Edwards, S.L. *et al.* Resistance to therapy caused by intragenic deletion in *BRCA2*. *Nature* **451**, 1111–1115 (2008).
42. Sakai, W. *et al.* Secondary mutations as a mechanism of cisplatin resistance in *BRCA2*-mutated cancers. *Nature* **451**, 1116–1120 (2008).

[1]Wellcome Trust Sanger Institute, Hinxton, UK. [2]Guy's and St Thomas' NHS Trust, London, UK. [3]Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Lund, Sweden. [4]Oncology, Innovative Medicines and Early Development Biotech Unit, AstraZeneca, Little Chesterford, UK. [5]Translational Research Lab Department, Centre Léon Bérard, Lyon, France. [6]Department of Medical Oncology, Erasmus MC Cancer Institute and Cancer Genomics, Erasmus University Medical Center, Rotterdam, the Netherlands. [7]Centre for Clinical Research and School of Medicine, The University of Queensland, Brisbane, Queensland, Australia. [8]Memorial Sloan Kettering Cancer Center, New York, New York, USA. [9]Cancer Research Laboratory, Faculty of Medicine, University of Iceland, Reykjavik, Iceland. [10]Department of Pathology, College of Medicine, Hanyang University, Seoul, Republic of Korea. [11]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK. [12]Department of Molecular Biology, Faculties of Science and Medicine, Radboud University, Nijmegen, the Netherlands. [13]Department of Pathology, Academic Medical Center, Amsterdam, the Netherlands. [14]Department of Cancer Genetics, Institute for Cancer Research, Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway. [15]K.G. Jebsen Centre for Breast Cancer Research, Institute for Clinical Medicine, University of Oslo, Oslo, Norway. [16]Department of Radiation Oncology, Radboud University Medical Center, Nijmegen, the Netherlands. [17]Department of Laboratory Medicine, Radboud University Medical Center, Nijmegen, the Netherlands. [18]Pathology Queensland, Royal Brisbane and Women's Hospital, Brisbane, Queensland, Australia. [19]Department of Pathology, Institut Curie, Paris, France. [20]INSERM U934, Institut Curie, Paris, France. [21]Breast Cancer Translational Research Laboratory, Université Libre de Bruxelles, Institut Jules Bordet, Brussels, Belgium. [22]Breast Cancer Now Research Unit, King's College, London, UK. [23]Breast Cancer Now Toby Robin's Research Centre, Institute of Cancer Research, London, UK. [24]Department of Breast Surgical Oncology, University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [25]Translational Cancer Research Unit, Center for Oncological Research, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium. [26]HistoGeneX, Wilrijk, Belgium. [27]Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts, USA. [28]Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [29]Equipe Erable, INRIA Grenoble-Rhône-Alpes, Montbonnot-Saint Martin, France. [30]Synergie Lyon Cancer, Centre Léon Bérard, Lyon, France. [31]East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [32]These authors contributed equally to this work. Correspondence should be addressed to S.N.-Z. (snz@sanger.ac.uk).

## ONLINE METHODS

**Data set.** The internal review boards of each participating institution approved collection and use of samples of all patients in this study. Informed consent was obtained by the relevant participating institution.

DNA was extracted from 560 individuals (556 females and 4 males) with breast cancer from both tumor and corresponding normal tissue and samples were subjected to whole-genome sequencing as described previously[27]. Resulting BAM files were aligned to the reference human genome (GRCh37) using Burrows–Wheeler aligner, BWA (v0.5.9)[43].

Mutation calling was performed as described previously[27]. Briefly, CaVEMan (Cancer Variants Through Expectation Maximization; http://cancerit.github.io/CaVEMan/) was used to call somatic substitutions. Indels in the tumor and normal genomes were called using modified Pindel version 2.0 (http://cancerit.github.io/cgpPindel/) on the NCBI37 genome build[44]. Structural variants were discovered using a bespoke algorithm, BRASS (BReakpoint AnalySiS) (https://github.com/cancerit/BRASS) through discordantly mapping paired-end reads followed by *de novo* local assembly using Velvet[45] to determine exact coordinates and features of breakpoint junction sequence.

In total, 3,479,652 somatic base substitutions, 371,993 small indels and 77,695 rearrangements were detected in the 560 samples[27].

*Mutational signatures background.* Mutation signature analysis based on non-negative matrix factorization (NMF) was performed as described previously[27,46]. Twelve consensus base-substitution signatures were identified previously in the 560 breast whole genomes: signatures 1, 2, 3, 5, 6, 8, 13, 17, 18, 20, 26 and 30.

Base-substitution signatures 1 (characterized by C>T transitions at N$\underline{C}$G, where the underlined base is mutated) and 5 (primarily characterized by C>T and T>C mutations) have previously been associated with age[47]. Both signature 2, which is predominantly composed of C>T substitutions at T$\underline{C}$N, and signature 13, which predominantly comprises C>G mutations at T$\underline{C}$N, may be generated by members of the AID/APOBEC family of cytidine deaminases that deaminate cytosine to uracil. Signatures 3 (enriched in C>G substitutions) and 8 (enriched in C>A substitutions) both lack highly distinctive substitution features and are enriched in *BRCA1*- and *BRCA2*-null tumors. Signature 3 in particular has been associated with the presence of inactivating *BRCA1* and *BRCA2* mutations. Signatures 6, 20 and 26 are associated with defective DNA mismatch repair and were restricted to 10 samples, which exhibited mutation profiles consistent with mismatch repair deficiency. The etiology of signatures 17 (characterized by T>G mutations at N$\underline{T}$T) and 18 (exhibiting a high proportion of C>A mutations) is unknown. Signature 30 (characterized by C>T transitions) was found in a single patient and may be due to previous exposure to cancer therapies.

Six rearrangement mutational signatures (RS1–RS6) based on rearrangement type, clustering and size were identified[27]. RS1 and RS3 were characterized by nonclustered tandem duplications. RS1 had mostly tandem duplications of >100 kb, while RS3 was predominantly associated with small tandem duplications of <10 kb. RS2 was characterized by nonclustered deletions (>100 kb), inversions and interchromosomal translocations. RS4 was characterized by clustered interchromosomal translocations. RS5 was associated with nonclustered deletions of <100 kb, while RS6 contained clustered inversions and deletions. RS5 is enriched in *BRCA1*/*BRCA2*-null tumors, and an excess of RS3 is observed in *BRCA1*-null tumors[27].

Two indel signatures, based on the presence of either short tandem repeats or short stretches of identical sequence at the breakpoints (termed overlapping microhomology), were also extracted. Deletions with microhomology were typically >3 bp in length and are characteristic of defective nonhomologous-end-joining-based DNA double-strand break repair, whereas indels at short tandem repeats are typical of the microsatellite instability associated with defective DNA mismatch repair. See **Supplementary Table 1** for the breakdown of the contribution of each mutation signature per sample.

*Determining whether a sample had particular mutational signatures.* An iterative algorithm was used to identify the set of COSMIC signatures (http://cancer.sanger.ac.uk/cosmic/signatures) active in each sample (the so-called 'exposure'). Each sample was completely described by a vector containing the number of substitutions observed for each mutation and flanking sequence context (defined by the neighboring bases immediately 5′ and 3′ to the mutated base and by the mutated base itself). Each mutation was oriented with respect to the pyrimidine strand, and, consequently, each vector contained 96 elements. The algorithm started from an initial solution estimated by using a simulated annealing-based method. Then, mutations were iteratively reassigned to alternative signatures, and cosine similarities were obtained by comparing the reconstructed 96-element vector for each potential reassignment to that of the observed 96-element vector with the aim of identifying the highest possible cosine similarity value that was possible. The algorithm stopped when no improvement to the cosine similarity was found.

*HRD indices.* Single-nucleotide polymorphism (SNP) array hybridization using the Affymetrix SNP6.0 platform was performed according to Affymetrix protocols. Allele-specific copy number analysis of tumors was performed using ASCAT (v2.1.1) to generate integral allele-specific copy number profiles for the tumor cells[48]. ASCAT was also applied to next-generation sequencing data directly, producing highly comparable results. Resulting allele-specific data generated by ASCAT were used in the calculation of the homologous recombination deficiency (HRD) index score using implementations made in R[49,50]. See **Supplementary Table 1** for HRD index scores.

*Variants in BRCA1 and BRCA2 and other HR genes.* For details of the process used to discover germline and somatic mutations in *BRCA1*/*BRCA2* and other genes known to be involved in DNA repair via homologous recombination, see the **Supplementary Note**.

**Lasso logistic regression modeling.** *Learning phase.* We set out to create a method for detecting genomic features associated with deficiency in *BRCA1*/*BRCA2* that would report the probability of a tumor sample being HR deficient during its evolution.

The method is trained on whole-genome sequencing data. We utilized the information on the signatures of single base substitutions, indels and rearrangements and a copy number classification based on HRD indices.

This supervised learning method was first applied to a cohort of 22 carriers of germline *BRCA1* and *BRCA2* mutations with clear loss of the other parental allele. There were also 235 control samples that did not have *BRCA1*/*BRCA2* mutations or promoter hypermethylation of *BRCA1*, or any evidence of signatures of *BRCA1*/*BRCA2* deficiency, and were believed to be sporadic breast cancers. *BRCA1*/*BRCA2*-proficient tumors are usually reported as relatively stable genomically and quiescent in mutational profile. Thus, using this prior knowledge, we manually interrogated genome plots of the overall mutation patterns to identify the 235 samples that we could confidently call *BRCA*-proficient tumors.

Inputs into the algorithm were as follows: (i) counts of mutations associated with each signature of single-base substitutions: signatures 1, 2, 3, 5, 6, 8, 13, 17, 18, 20 and 26 (signature 30 was excluded as it involved only 1 sample), (ii) indels with microhomology at the indel breakpoint junction, indels at polynucleotide-repeat tracts and other complex indels as proportions, (iii) counts of rearrangements associated with each signature of rearrangements RS1–RS6 and (iv) HRD index.

Some samples had vastly higher counts of substitutions than others, and such outliers posed a challenge in the analysis. Thus, the genomic features were first log transformed, according to the following formula.

$$x' = \ln(x + 1) \tag{1}$$

The ranges of values for each class of mutation were vastly different. Therefore, the transformed data were normalized so that each feature had a mean of 0 and s.d. of 1, in order to be able to make the features comparable to one another.

$$x'' = \frac{x' - \text{mean}(x')}{\text{s.d.}(x')} \tag{2}$$

A lasso logistic regression[51] model was used to identify the genomic features that could distinguish the two categories of patient samples: those affected and those not affected by *BRCA1*/*BRCA2* deficiency. An efficient computer implementation for learning model parameters was available through the R package glmnet. The lasso approach permits learning and weighting of the genomic features most relevant to predicting *BRCA1*/*BRCA2* status through variable selection.

$\beta$ coefficients, also referred to as weights, are learned from the genomic features of *BRCA1*/*BRCA2*-proficient and *BRCA1*/*BRCA2*-deficient samples presented to the algorithm. Optimal coefficients are obtained by minimizing the objective function[51]

$$\min_{((\beta 0,\beta))\in\mathbb{R}^{p+1}} \left( -\left[ \frac{1}{N} \sum_{i=1}^{N} y_i \cdot (\beta_0 + x_i^T \beta) - \log\left( 1 + e^{(\beta_0 + x_i^T \beta)} \right) \right] + \lambda \, || \beta ||_1 \right) \quad (3)$$

where $y_i$ is the *BRCA* status of a sample: $y_i = 1$ for *BRCA1*/*BRCA2*-null samples, $y_i = 0$ otherwise; $\beta_0$ is the intercept, interpreted as the log of odds of $y_i = 1$ when $x_i{}^T = 0$; $\beta$ is a vector of weights, each corresponding to a genomic feature; $p$ is the number of features characterizing each sample; $N$ is the number of samples; $x_i^T$ is the vector of features characterizing the $i$th sample; $\lambda$ is the penalty promoting the sparseness of the weights, as learned through nested cross-validation; and $||\beta||_1$ is the L1 norm of the vector of weights (i.e., the sum of the absolute values of all entries of the coefficient vector)

We constrained all $\beta$ weights to be positive because they reflect the biological presence of mutational processes that are due to (in this case) *BRCA1*/*BRCA2* deficiency. By setting the constraint of nonnegative weights, we ensured that all samples would be scored on the basis of the presence of relevant mutational signatures associated with *BRCA1*/*BRCA2* deficiency, irrespective of whether these signatures are the dominant mutational process in the cancer.

Multiple mutational processes can exist in a tumor, and, in some cases, certain hypermutator mutational phenotypes can come to dominate a specific cancer and eclipse the appreciation of other mutational processes. However, using nonnegative coefficients in our model ensures that mutational signatures associated with *BRCA1*/*BRCA2* deficiency are detected reliably no matter how weakly they are present.

Ultimately, the lasso logistic regression model is used to assign a probabilistic score to any new sample that is being analyzed, using the normalized exposures of mutational processes in the sample ($x_i^T$) and applying the parameters of the model ($\beta$) as follows

$$P(C_i = BRCA) = \frac{1}{1 + e^{-(\beta_0 + x_i^T \beta)}} \quad (4)$$

where $C_i$ is the variable encoding the status of the $i$th sample; $\beta_0$ is the intercept weight; $x_i^T$ is the vector encoding features of the $i$th sample; and $\beta$ is the vector of weights.

*Robustness, stability and generalizability (on 22).* We trained the logistic regression model using a cohort of 22 germline *BRCA1*/*BRCA2* mutation carriers with loss of the second allele and a cohort of 235 sporadic tumors in this supervised analysis.

We used a tenfold nested cross-validation strategy to assess the robustness and generalizability of the learned weights. Ten outer folds were used in the cross-validation process where 10% of data were set aside for each outer fold and were used to assess the accuracy of the prediction and generalizability.

The remaining 90% of the data were used for model parameter selection. The parameters associated with *BRCA1*/*BRCA2* deficiency were investigated on the inner folds for a range of $\lambda$ values that define the sparsity of the results.

We obtained the model coefficients across the ten folds (presented as box plots in **Supplementary Fig. 7**), demonstrating that the results across the ten folds are consistently nonzero for each of the genomic parameters identified as distinguishing. The model was finally applied across all the data used in training, with the coefficients from this final run (**Supplementary Table 10**) also presented as red crosses in **Supplementary Figure 7**.

The genomic parameters and associated coefficients were identified for a $\lambda$ value of 0.000480 (mean, 0.000891; s.d., 0.000803), to distinguish samples with HR deficiency in the final step.

Finally, we assessed the stability of each coefficient through subsampling of the training set. We chose half of samples in the training set randomly and counted how many times each genomic feature was selected as a distinguishing feature (i.e., had a nonzero coefficient). This was performed iteratively; out of 100 subsampling and training iterations, each coefficient was nonzero (**Supplementary Table 11**).

While most features are relatively stable, rearrangement signature RS3, which is a feature of *BRCA1*-null tumors, appears to be less stable. This lower stability is likely due to the cohort of 22 informative tumors that were chosen to represent *BRCA1*/*BRCA2* nullness. Only 5 of the 22 patients are *BRCA1*-null patients; thus, there is a skew in the cohort to *BRCA2*, and the balance in this cohort could be improved as a means to improve the learned weights and their relative stability.

*Identifying further samples with BRCA1/BRCA2 deficiency.* The logistic regression model was applied to all 560 samples in the cohort. In particular, we could calculate the BRCAness scores on samples that were not in the training set, for example, because their genomes were not quiescent or had uncertain genomic profiles. **Supplementary Figure 2** shows the scores for each sample, demonstrating a steep sigmoidal curve.

Apart from the 22 individuals recruited into the study, many individuals with high BRCAness scores had germline mutations that we had not known of at the time of their enrollment into the study, and many had somatic *BRCA1*/*BRCA2* mutations. All had loss of the other parental allele. We thus reasoned that features of BRCAness are present in samples with biallelic inactivation of *BRCA1* and *BRCA2* genes, whether germline or somatic, and included all such samples in a further round of training.

*Retraining on 77 samples and defining HRDetect, a classifier of BRCA1/BRCA2 deficiency.* In the final round of training, we included 77 samples with biallelic inactivation of *BRCA1*/*BRCA2* and 234 quiescent tumors as negative examples. The number of *BRCA1*/*BRCA2*-proficient tumors differs by one sample between this and the previous training round because one of the samples with a quiescent genome, PD6042a, was subsequently found to have a biallelic mutation of *BRCA2*. We assessed the robustness and generalizability of HRDetect using nested cross-validation as before (see Online Methods, "Robustness, stability and generalizability (on 22)").

The genomic parameters and associated coefficients learned across the ten folds of cross-validation are shown in **Supplementary Figure 8**. The box plots show the variability of each coefficient, and the red crosses show the values of the coefficients when training on the whole data set.

In comparison to training with only 22 known carriers of germline *BRCA1*/*BRCA2* mutations, the variability of the coefficient values across folds was decreased. A larger number of informative samples in the training set improved the robustness of the coefficients.

With the higher number of training samples, the stability of individual coefficients also improved, as shown in **Supplementary Table 12**.

The logistic regression model was settled on the coefficients in **Supplementary Table 13**, with $\lambda$ of 0.00369 (mean, 0.00478; s.d., 0.00104).

The accuracy of the BRCA predictions was excellent, with an area under the curve in cross-validation of 1 for the 77 *BRCA1*/*BRCA2*-null samples and 234 quiescent tumors (311 samples from the 560 breast cancer genomes).

We also explored the possibility of permitting interactions between all genomic covariates in order to discover potentially augmented effects from cooperating signatures in our model (for details, see the **Supplementary Note**).

*Assessment of the accuracy of the classifiers through ROC curves.* For a more comprehensive assessment of the accuracy of HRDetect, we extended the set of samples by 60 samples that had been excluded from training. We applied HRDetect to all samples that had been successfully characterized with respect to methylation and HRD indices. We ultimately assessed the performance of HRDetect on 371 of the 560 breast cancer genomes, ignoring 2 samples with no HRD index and 187 samples with missing methylation data, as their *BRCA1*/*BRCA2* status could not be verified.

In calculating the ROC curves, we compared the predictions from HRDetect for each of the 371 samples against evidence of biallelic loss of *BRCA1*/*BRCA2*. The area under the ROC curve for the breast cancer genomes was 0.98.

Finally, HRDetect was applied to the full set of 560 breast cancer samples to give the final HRDetect score for each sample.

The flow diagram in **Supplementary Figure 9** describes the steps involved in training, evaluation and application of HRDetect to the full data set.

*Applying HRDetect to new tumor samples.* Applying the predictor to a new sample requires characterization of the sample with respect to the signatures of single-base substitutions, rearrangements, copy number profile and HRD score, and small insertions and deletions together with the characteristics of adjacent sequence.

Furthermore, the features of a new sample need to be normalized as in equations (1) and (2). See **Supplementary Table 14** for the means and s.d. of each feature that were taken into account based on current settings of HRDetect.

After the features of a new sample were normalized, the HRDetect score was obtained by applying equation (4), coefficients from **Supplementary Table 13** and intercept $\beta 0 = -3.364$.

For details of how HRDetect was applied to new breast cancer samples, downsampled genomes, and breast cancer WES and WGS data from other cancer types, see the **Supplementary Note**.

**Data availability.** Breast cancer whole-genome sequence BAM files and CEL files from Affymetrix SNP6 arrays are available from the European Genome-phenome Archive (EGA).

The overarching EGA accession number for the 560 breast cancers used in the initial development of HRDetect is EGAS00001001178. This includes both whole-genome sequence BAM files and SNP6 array CEL files.

The accession numbers for the 80 additional breast cancers used for validation are EGAD00001002740 (sequence BAM files) and EGAD00010001079 (SNP6 array CEL files).


43. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
44. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
45. Zerbino, D.R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
46. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J. & Stratton, M.R. Deciphering signatures of mutational processes operative in human cancer. *Cell Reports* **3**, 246–259 (2013).
47. Alexandrov, L.B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
48. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. USA* **107**, 16910–16915 (2010).
49. Abkevich, V. *et al.* Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br. J. Cancer* **107**, 1776–1782 (2012).
50. Natrajan, R. *et al.* Characterization of the genomic features and expressed fusion genes in micropapillary carcinomas of the breast. *J. Pathol.* **232**, 553–565 (2014).
51. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 267–288 (1996).