HRDetect: A mutational signature based predictor of *BRCA1* and *BRCA2* deficiency

Authors

*Helen Davies[1], *Dominik Glodzik[1], 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], Sunil R Lakhani[7,17], Anne Vincent-Salomon[18], Christos Sotiriou[19], Andrew Tutt[20,21], Alastair M. Thompson[22], Steven Van Laere[23,24], Andrea L. Richardson[25,26], Alain Viari[27,28], Peter J Campbell[1], Michael R. Stratton[1] and Serena Nik-Zainal[1,29]

[AU: please check author spelling and affiliations]
*shared first authorship

Affiliations

1 Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK
2 Guys and St Thomas' NHS Trust, London, UK
3 Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Lund, SE-223 81, Sweden
4 Oncology, Innovative Medicines and Early Development Biotech Unit, AstraZeneca, Hodgkin Building, Chesterford Research Park, Little Chesterford, Cambridge CB10 1XL, UK
5 Centre Léon Bérard, Translational Research Lab Department, 28, rue Laënnec, 69373 Lyon Cedex 08, France
6 Department of Medical Oncology, Erasmus MC Cancer Institute and Cancer Genomics Netherlands, Erasmus University Medical Center, Rotterdam 3015CN, The Netherlands
7 The University of Queensland: UQ Centre for Clinical Research and School of Medicine, Brisbane, Queensland 4029, Australia
8 Memorial Sloan Kettering Cancer Center, 1275 York Ave, New York, NY 10065, United States
9 Cancer Research Laboratory, Faculty of Medicine, University of Iceland, 101 Reykjavik, Iceland
10 Department of Pathology, College of Medicine, Hanyang University, Seoul, 133-791, South Korea
11 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus,Hinxton, Cambridgeshire, CB10 1SD
12 Department of Molecular Biology, Faculties of Science and Medicine, Radboud University, 6525GA, Nijmegen, Netherlands
13 Department of Pathology, Academic Medical Center, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands

14 Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital The Norwegian Radium Hospital Oslo 0310, Norway
15 K.G. Jebsen Centre for Breast Cancer Research, Institute for Clinical Medicine, University of Oslo, Oslo 0310, Norway
16 Department of Radiation Oncology, and department of Laboratory Medicine, Radboud university medical center, Nijmegen 6525GA, The Netherlands.
17 Pathology Queensland, The Royal Brisbane and Women's Hospital, Brisbane, Queensland 4029, Australia
18 Institut Curie, Department of Pathology and INSERM U934, 26 rue d'Ulm, 75248 Paris Cedex 05, France
19 Breast Cancer Translational Research Laboratory, Université Libre de Bruxelles, Institut Jules Bordet, Bd de Waterloo 121, B-1000 Brussels, Belgium
20 Breast Cancer Now Research Unit, King's College, London, UK
21 Breast Cancer Now Toby Robin's Research Centre, Institute of Cancer Research, London, UK
22 Department of Breast Surgical Oncology, University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, Texas 77030, USA
23 Translational Cancer Research Unit, Center for Oncological Research, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium
24 HistoGeneX NV, Wilrijk, Belgium
25 Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115 USA
26 Dana-Farber Cancer Institute, Boston, MA 02215 USA
27 Equipe Erable, INRIA Grenoble-Rhône-Alpes, 655, Avenue de l'Europe, 38330 Montbonnot-Saint Martin, France
28 Synergie Lyon Cancer, Centre Léon Bérard, 28 rue Laënnec, Lyon Cedex 08, France
29 East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge CB2 9NB, UK

Corresponding author : Serena Nik-Zainal (snz@sanger.ac.uk)

Abstract

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**.** Germline and/or somatic mutations in *BRCA1*/*BRCA2* in other cancer types also confer selective sensitivity to PARP inhibitors. Thus, assays to detect *BRCA1*/*BRCA2* deficient tumours have been sought. Recently, somatic substitution, insertion/deletion and rearrangement patterns or *mutational signatures* were associated with *BRCA1*/*BRCA2* dysfunction. We used a supervised lasso logistic regression model to identify six critically 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 tumours with 98.7% sensitivity (AUC 0.98). Application of this model in a cohort of 560 breast cancer patients with 22 known germline *BRCA1*/*BRCA2* mutation carriers, allowed us to identify an additional 22 somatic *BRCA1*/*BRCA2* null tumours and 47 tumours with functional *BRCA1*/*BRCA2*-deficiency where no mutation was detected. We validated HRDetect on independent cohorts of breast, ovarian and pancreatic cancers, and demonstrate efficacy on alternative sequencing strategies. Integrating all classes of mutational signatures thus reveals a larger proportion of breast cancer patients (of up to 22%) than hitherto appreciated (~1-5%) that could have selective therapeutic sensitivity to PARP-inhibition.

Introduction

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 tumour suppressor genes usually lose the wild-type allele during tumorigenesis to become fully inactivated[7]. *BRCA1*/*BRCA2* null tumours are thus deficient in HR and selectively sensitive to compounds that increase the demand on HR[8]. Poly (ADP-ribose) polymerase (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 tumour cell death[10,11].

Preclinical studies and Phase I/II breast and ovarian clinical trials[12,13] have shown PARP-inhibitor efficacy in familial *BRCA1* and *BRCA2* patients. However, PARP-inhibition has applications beyond that of germline mutated tumours[14]. Effective PARP-inhibition maintenance therapy has been demonstrated in high grade serous ovarian cancer with germline or somatic *BRCA1*/*BRCA2* mutations[15]. Thus, extensive efforts have been put into identification of molecular features of tumours that are *BRCA1*/*BRCA2* deficient, referred to historically as "BRCAness", whether inactivated through germline, somatic or secondary means, including promoter 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], transcriptional metagene signatures[18-20], copy number-based methods (e.g. HRD (Homologous Recombination Deficiency) index and genomic "scars")[21-23] and functional assays of HR competence[24] have been developed to detect *BRCA1*/*BRCA2* deficiency. However, these indices have had limited predictive success. A recent review suggests that a good predictor of the biological status of a HR-deficient tumour is essential, as the cohort of tumours that demonstrate BRCAness and that could be selectively sensitive to PARP-inhibitors is likely not limited to the small proportion of familial breast and ovarian cancers, 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 significantly 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 on-going exposures, whether of environmental insults such as ultraviolet radiation, or of endogenous biochemical degradation and deficiencies of DNA repair pathways like HR.

We reason that mutational signatures which report *BRCA1*/*BRCA2* deficiency in germline mutated tumours could be used as a predictor of other tumours that also have this deficiency. Previously, base substitution Signature 3 was shown to distinguish germline *BRCA1/BRCA2* null from sporadic cancers in a small subset of breast cancers[29,30] and subsequently extended to pancreatic[32,33], ovarian[34] and stomach cancer[35]. However, selecting a cut-off to discriminate *BRCA1*/*BRCA2*-deficient from -proficient cancers is not straightforward when using this signature alone. Recent characterisation of a large cohort of WGS breast cancers[27,28] has provided new insights. A defect in a single gene such as *BRCA1*/*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 genomic features of *BRCA1*/*BRCA2* deficiency and present a WGS-based predictor with remarkable preformance for detection of HR-deficient tumours.

Results

Quantitatively defining features of "BRCA"ness

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

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

The 22 *BRCA1*/*BRCA2* null tumours were used in a first training set to quantitatively define features of *BRCA1*/*BRCA2* deficiency. They were contrasted to a cohort of 235 sporadic breast cancers with quiescent genomic profiles, distinct from *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 supervised learning lasso logistic regression model was applied to counts of mutational signatures and to HRD indices that were log-transformed and normalized to permit comparability between genomic parameters (Supplementary Table 2).

An iterative ten-fold 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 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 similar characteristics to germline *BRCA1*/*BRCA2* null tumours Figure 2 for workflow, Supplementary Table 2). The resulting distribution of probabilities of *BRCA1*/*BRCA2* deficiency was a strikingly steep sigmoidal curve with clear distinction between patients predicted to have high or low probabilities of *BRCA1*/*BRCA2* deficiency. Apart from the 22 positive controls from the training set, 90/538 additional tumor samples were identified as having a probability of *BRCA1*/*BRCA2* deficiency exceeding 70%, 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/BRCA2* mutations (germline and somatic) in the cohort of 560 patients. 33 were found to carry pathogenic germline variants in *BRCA1*/*BRCA2* with corresponding somatic inactivation of the alternative allele. This more than doubles the number of women harbouring familial cancer predisposition alleles than originally recruited into the study, carrying significant clinical genetic counselling implications and potential for active surveillance and/or treatment choices, for affected patients and their families.

22 patients had early, clonal, somatic *BRCA1*/*BRCA2* mutations (eight) or promoter DNA hypermethylation of *BRCA1* (fourteen) with inactivation of the alternative allele. The remaining tumours with probability of BRCA1/BRCA2 deficiency exceeding 70% did not demonstrate biallelic inactivation of *BRCA1*/*BRCA2,* although DNA methylation data were not available for a subset of patients.

6 BRCA1-null samples had probabilities of 0.006-0.64 and were missed because the algorithm had been trained on a small cohort of five *BRCA1* tumours out 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 patients were identified as null for *BRCA1*/*BRCA2*, we performed another iteration of the lasso logistic regression model on a larger, better-powered training set comprising 77 samples (22 known germline, 33 new germline diagnoses, 22 somatic)(Figure 2).

Reassuringly, the same genomic features were identified as predictive parameters as was observed for the 22 germline null samples, with the addition of base substitution Signature 8, ranked by decreasing weight – microhomology-mediated deletions(2.398), base substitution signature 3(1.611), rearrangement signature 3(1.153), rearrangement signature 5(0.847), HRD index(0.667) and base substitution signature 8 (0.091)(Figure 3, Supplementary Table 3). Acknowledging the imbalance in numbers of *BRCA1*/*BRCA2* null cancers versus controls, supervised learning was repeated in a 1:1 ratio. Differences to the original assessment (77 null : 234 control) were insignificant, verifying the stability of the six critically distinguishing parameters. With a larger dataset available, we permitted identification of interactions between genomic covariates in order to discover potentially augmented effects of cooperating signatures. Although correlations were observed, the performance including interactions did not improve on predictions when compared to the model without interactions. We therefore opted for the simpler model, keeping each genomic parameter independent. Thus, we finalised our predictor of *BRCA1*/*BRCA2* deficiency, termed HRDetect, on this set.

HRDetect was re-applied on the cohort of 560 breast cancers and showed excellent performance as revealed by a receiver operating characteristic (ROC) curve with an area under the ROC curve (AUC) of 0.98 (Figure 4). This result is unlikely to be bettered, emphasising the value of utilising multiple mutational signatures as a read-out of *BRCA1*/*BRCA2* deficiency. Reinforcing this point, no individual genomic parameter performed as well as all six genomic signatures incorporated together in HRDetect (Figure 4). In particular, it is superior to current methods of assessing *BRCA1*/*BRCA2* deficiency, specifically, the genomic "scar" based index[21-23] like the HRD score (sensitivity of HRD score = 60%, Figure 4 ROC curve compares HRDetect against HRD score alone and other mutational signatures individually).

Using a probabilistic cut-off of 70%, HRDetect predicts *BRCA1*/*BRCA2* deficiency with a sensitivity of 98.7% in the cohort of 560 patients. HRDetect reveals a total of 124 samples with a score exceeding 70%, including an additional 47 samples with a high probability of *BRCA1*/*BRCA2* deficiency. These remaining tumours (5/340 ER positive and 42/143 ER negative) with high scores and neither germline nor somatic *BRCA1*/*BRCA2* mutations and for which promoter hypermethylation of *BRCA1* was either not observed (10) or not available for assessment (37), were investigated for inactivation of other genes involved in HR repair and for other germline susceptibility alleles.

Other genetic factors and "BRCA"ness

Of these 47 samples, three had mutations in HR genes. One patient, PD4875a, exhibited a high HRDetect score (0.94) and showed a profile 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 tumour. Inactivation of the wild-type allele by alternative means cannot be excluded. This patient also carried a germline truncating mutation in *BRIP*1 (a gene associated with moderate penetrance breast cancer risk) with loss of the alternative allele; however with only a single example of a truncating BRIP1, its significance is unclear. Otherwise, all other tumours with monoallelic germline/somatic inactivation of *BRCA1*/*BRCA2* were associated with low HRdetect scores (thirteen patients total: four germline, seven somatic, two *BRCA1* promoter hypermethylation) (Supplementary Table 4).

Two patients (PD24205a and PD24212a) had somatic monoallelic *PALB2* (a gene associated with moderate penetrance breast cancer risk) truncating mutations. A third patient which had low HRDetect score, PD11340a, also had a deleterious somatic monoalleleic essential splice *PALB2* mutation. Given the small numbers, we would interpret the contribution of *PALB2* mutations with caution as other modes of *BRCA1*/*BRCA2* inactivation or 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 from the list of HR genes (*RAD51C*, *RAD50, CHEK2* and *FANCA-FANCN)* was identified as a contributor amongst tumours with high scores. Notably, high- and moderate-penetrance germline breast cancer susceptibility alleles including *TP53, PTEN, ATM*, *CHEK2*, *ATR*, *RAD50*, *CDH1, STK11, PALB2,* whether remaining monoallelic or demonstrating biallelic inactivation, were neither associated with a genomic profile nor a high probability of *BRCA1*/*BRCA2* deficiency.

These results firstly, emphasise the importance of knowing the status of the alternative parental allele in the interpretation of mutation data (Supplementary Table 4). Secondly, they also highlight that nearly a third of tumours with high scores for *BRCA1*/*BRCA2* deficiency scores cannot be authenticated as BRCA1/BRCA2 null through genetic/epigenetic means. Yet, given the striking resemblance to BRCA1/BRCA2 null tumours, it is intriguing to consider that these cancers are therefore biologically comparable and likely to respond similarly to BRCA1/BRCA2 null cancers, in particular to PARP-inhibition.

Validation of HRDetect in a new cohort of 80 WGS breast cancer

HRDetect was applied to a new cohort of WGS breast cancers from 80 women (Figure 4, Supplementary Table 5) with mainly ER positive, HER2 negative breast cancer as a validation exercise. HRDetect successfully identified 1 germline *BRCA1* and 5 *BRCA2* mutation carriers (4 germline and 1 somatic) with associated loss of the wild type allele. One sample, PD14434a, carried a germline essential splice site mutation with loss of the alternative parental allele but fell short of the HRDetect cut off of 70%. Two samples, one germline *BRCA2* and one 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 on alternative sequencing strategies

To explore the performance of HRDetect on alternative sequencing strategies, we performed an *in silico* experiment, randomly down-sampling the sequences of the 560 high coverage (30- to 40-fold) WGS breast cancers, to generate low coverage (10-fold, range 9.9 to 10.5) WGS sequence files for analysis (Supplementary Table 6). Somatic mutations were re-called across down-sampled 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 representative of over-arching mutation patterns are present.

As expected, the numbers of base substitutions, indels and rearrangements were consistently lower in the down-sampled *in silico* experiment of all samples when compared to the original high-coverage experiment. Nevertheless, all twelve base substitution, two indel and six rearrangement signatures were detectable, at approximately the same proportions per sample, albeit at reduced absolute numbers. Additionally, copy number analysis showed good concordance for overall HRD scores (r=0.63) between high and low coverage genomes. Thus, despite the reduction in sensitivity of individual somatic mutations, the detection of over-arching mutation signatures remained relatively secure.

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

By contrast, when HRDetect is used to assess *BRCA1*/*BRCA2* deficiency on data that are representative of only coding sequences (whole exome sequencing, WES), the sensitivity of detection is affected considerably falling to 46.8%. This is because 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 Information). When the HRDetect algorithm is retrained taking WES-based data as input alone, the performance of the classifier is improved (sensitivity 73% (56/77)) 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 cancers

HRDetect was applied to other WGS cancers including pancreatic and ovarian cancers to assess generalizability across tumour types[32-34]. Available BAM files were recalled through our somatic-mutation calling pipeline, mutational signatures extracted and copy number profiles obtained.

The ovarian cancer cohort comprised 73 samples. Using a threshold of 70%, 46 (63%) were identified as having a high probability of *BRCA1*/*BRCA2* deficiency. Of these, 30 were confirmed as having germline or somatic *BRCA1*/*BRCA2* mutations with loss of the wild-type parental allele (germline (14), somatic (6) and DNA methylation (10))(Supplementary Table 8). None were missed. Thus, again HRDetect has a sensitivity of detecting *BRCA1*/*BRCA2* null cancers approaching 100% and has uncovered 16 additional patients 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*/*BRCA2* (three germline and two somatic) and had lost the wild-type allele, one had retained the other allele. Epigenetic data were not available to interrogate the status of the remaining five samples. Three samples had *BRCA2* mutations but did not demonstrate convincing evidence of loss of the alternative allele and had low HRDetect scores (Supplementary Table 8). Thus, HRDetect had a sensitivity approaching 100% in this pancreatic cancer cohort and identified 5 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 (Figure 4C). When more samples become available 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 and their relevance to accelerating clinical application

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

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 find that *BRCA1*/*BRCA2* deficient cancers remain consistently identified by HRDetect despite excessive APOBEC-related mutagenesis in some samples (Supp Fig 7 for example). Furthermore, HRDetect is able to discern *BRCA1*/*BRCA2* deficient cancers with remarkable precision over a wide range of tumour cellularities, including relatively low cellularity samples (but not less than 15%) where mutation-calling sensitivity may be compromised. Thus, irrespective of biological or non-biological noise, HRDetect faithfully detects the signal of *BRCA1*/*BRCA2* deficiency, reinforcing the exceptional utility of this classifier.

Finally, to advance 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 four post-operative tumour block specimens) from nine patients with triple negative tumours that were treated with neoadjuvant anthracyclines +/- taxanes[37] (Supplementary Table 9, Supplementary Information). Although a different compound from PARP inhibitors, sensitivity to anthracyclines has been reported for tumours 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 were sporadic tumours (Figure 5C). By contrast, five patients that exhibited residual disease had low HRDetect probability scores. Furthermore, HRDetect performed consistently in independent biopsies per patient, and between biopsy and post-operative specimen per patient, without exception. Though the numbers are small, in all, these analyses suggest that HRDetect has potential for distinguishing therapeutic sensitivity as early in the patient's clinical journey as the first biopsy, and is robust between biopsies/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 (VUS) including 20 common alleles and 107 rare or private variants were identified in the 560 breast cancer dataset (Supplementary Table 4). 56 had concurrent loss of the other allele. However, these did not consistently demonstrate high scores for *BRCA1*/*BRCA2* deficiency emphasising that these variants VUSs are unlikely to be pathogenic and hence of low clinical significance.

There was one exception – PD23563a had a missense p.L1780P mutation in *BRCA1* with LOH of the other allele and a high score for *BRCA1*/*BRCA2* deficiency. This variant remains "of uncertain significance" in clinical databases of *BRCA1*/*BRCA2* SNP alleles (http://www.ncbi.nlm.nih.gov/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, 8 missense somatic *BRCA1/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 tumours

Thus far, we have focused on detecting tumours with either *BRCA1* or *BRCA2* deficiency and not distinguishing between them. Currently there is no clinical indication to separately identify these tumours because *BRCA1* and *BRCA2* deficient tumours are both similarly sensitive to PARP inhibition. In the future, however, reasons for separating these tumours may arise. The discriminating genomic parameters that distinguish *BRCA1* from *BRCA2* tumours are rearrangement Signature 3 and deletions without distinctive junctional characteristics (Supplementary Information). Both are detected best using WGS approaches.

Discussion and conclusions

Abrogation of *BRCA1*/*BRCA2* leads to not one, but a characteristic set of mutational signatures. As a predictive tool, utilising these multiple pathognomonic mutational signatures is extraordinarily effective with performance metrics which suggest that this method cannot be easily bettered. It dependably detects *BRCA1*/*BRCA2* deficiency in the presence of biological noise (e.g. APOBEC-related mutagenesis), when there is relatively low tumour cellularity (Supplementary Information) and when there is non-biological noise from formalin fixation. Indeed, demonstrating HRDetect efficacy in FFPE-banked samples opens doors in terms of exploration of historic and/or existing clinical trials, assuming matched normal DNAs are also available. Our analyses also emphasise that a WGS approach (even if at low-coverage (10 fold)) is far more effective than a WES approach at detecting *BRCA1*/*BRCA2* deficiency. Additionally, distinguishing *BRCA1* from *BRCA2* tumours is dependent on WGS-based methods. These methodological points have implications in designing genomic aspects of clinical trials. At least for detecting *BRCA1*/*BRCA2* deficiency, WGS approaches are optimal. Indeed, our analyses provide support and context for large-scale national WGS endeavours such as the UK 100,000 genomes project (https://www.genomicsengland.co.uk/the-100000-genomes-project/) and the Precision Medicine Initiative (http://www.cancer.gov/research/key-initiatives/precision-medicine) in the USA.

While the performance of HRDetect is extremely promising, algorithmic developments are envisaged including identification of tumours that have developed resistance alleles[41,42]. Because historic scars of HR deficiency will be present in a resistant tumour, this may lead to high HRDetect scores. However, distinguishing on-going 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 tumour[30,37].

Although only 22 patients were originally recruited with known germline *BRCA1*/*BRCA2* null cancers, HRDetect reveals an additional 33 germline, 22 somatic and 47 tumours where no mutation was detected – bringing the total to 124 (22%) *BRCA1*/*BRCA2* deficient tumours. 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 tumour's biological status and hence possible sensitivity to PARP inhibition (or other treatments (e.g. platinum-based salts, anthracyclines and mitomycin C) that cancers with BRCAness are selectively sensitive to). Nearly a 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 *BRCA1*/*BRCA2* genes and performing methylation assays would miss this cohort of patients that could be functionally deficient incurred through currently unknown means.

If the tumours 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, currently reserved for just ~1-5% of breast cancer patients who are germline mutation carriers. This is potentially transformative and thus 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 tumour, 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.

Figures

Figure 1

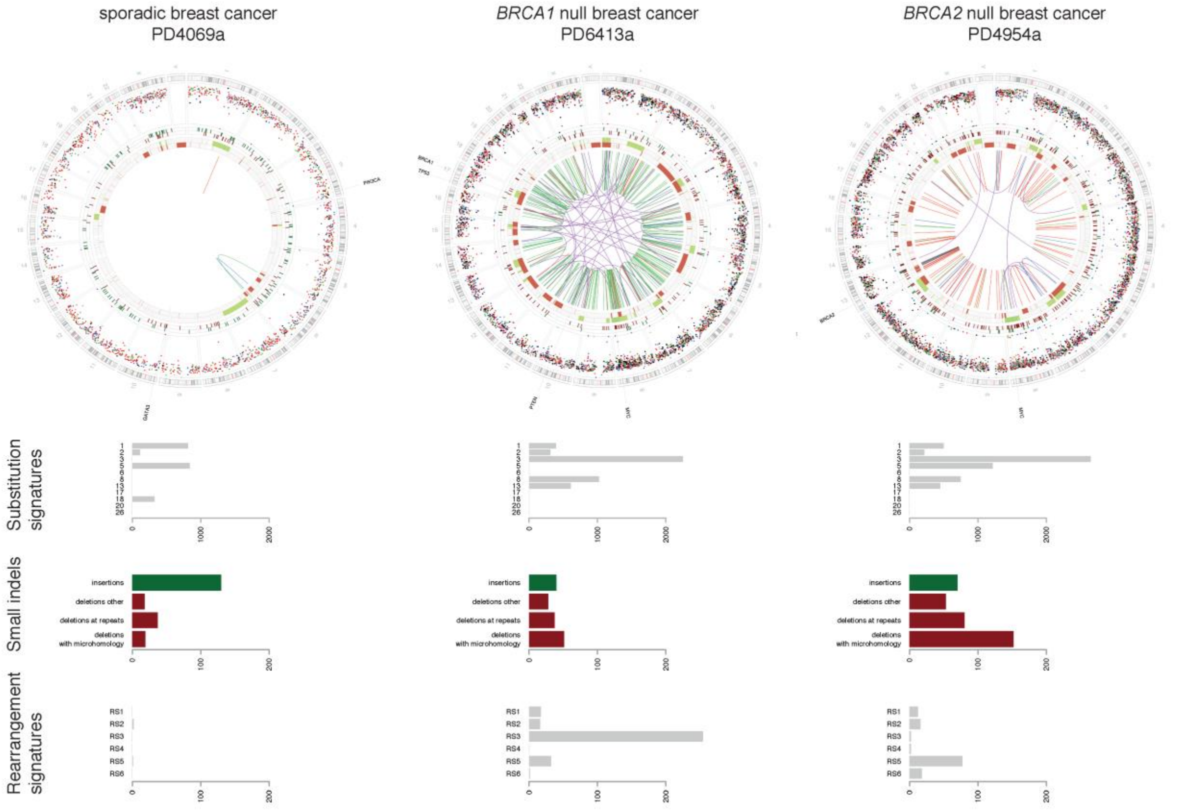


Figure 1: Whole genome profiling depicts differences between patients with *BRCA1*/*BRCA2* mutated tumours and sporadic tumours.

Examples of genome plots for a typical sporadic breast cancer (left), a *BRCA1* germline null (middle) and a *BRCA2* germline null tumour (right). Features depicted in circos plots from outermost rings heading inwards: Karyotypic ideogram outermost. Base substitutions next, plotted as rainfall plots (log10 intermutation distance on radial axis, dot colours: blue, C>A; black, C>G; red, C>T; grey, T>A; green, T>C; pink, T>G). Ring with short green lines, insertions; ring with short red lines, deletions. Major copy number allele (green, gain) ring, minor copy number allele ring (red, loss), Central lines represent rearrangements (green, tandem duplications; red, deletions; blue, inversions; purple, interchromosomal events). Mutations in breast cancer driver genes are indicated around the circus plots. Below each circos plot are the histograms showing mutation counts for each mutation class; topmost histogram shows the number of mutations contributing to each substitution signature; middle histogram represents indel patterns; lowermost histogram shows the number of rearrangements contributing to each rearrangement signature.

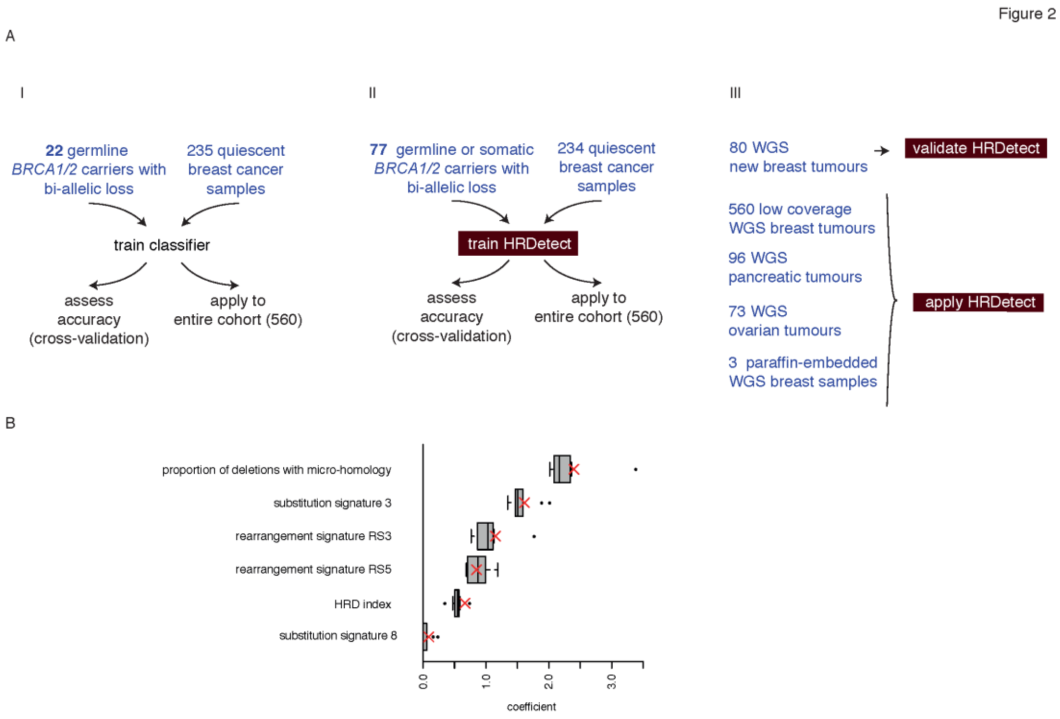


Figure 2: Workflow for developing HRDetect

A. Workflow of the steps involved in the development of the HRDetect predictor

I. Initial training using 22 known germline *BRCA1* and *BRCA2* null samples.

II. Retraining using 77 *BRCA1* and *BRCA2* null samples to produce the final HRDetect predictor.

III. Validation on a further set of breast cancers and application to other data sets.

The number of BRCA1/BRCA2 proficient tumours differs by one sample between the two training rounds because one of the samples with a quiescent genome, PD6042a, was subsequently found to have a biallelic mutation of *BRCA2*. **[**

B. Box plots of the weights for the genomic features contributing to the HRDetect predictor. Range of values from 10 replicates of training in cross-validation, using 311 breast cancer samples; 77 BRCA1/BRCA2 null samples and 234 quiescent tumours. Red crosses indicate the final coefficients used in HRDetect.

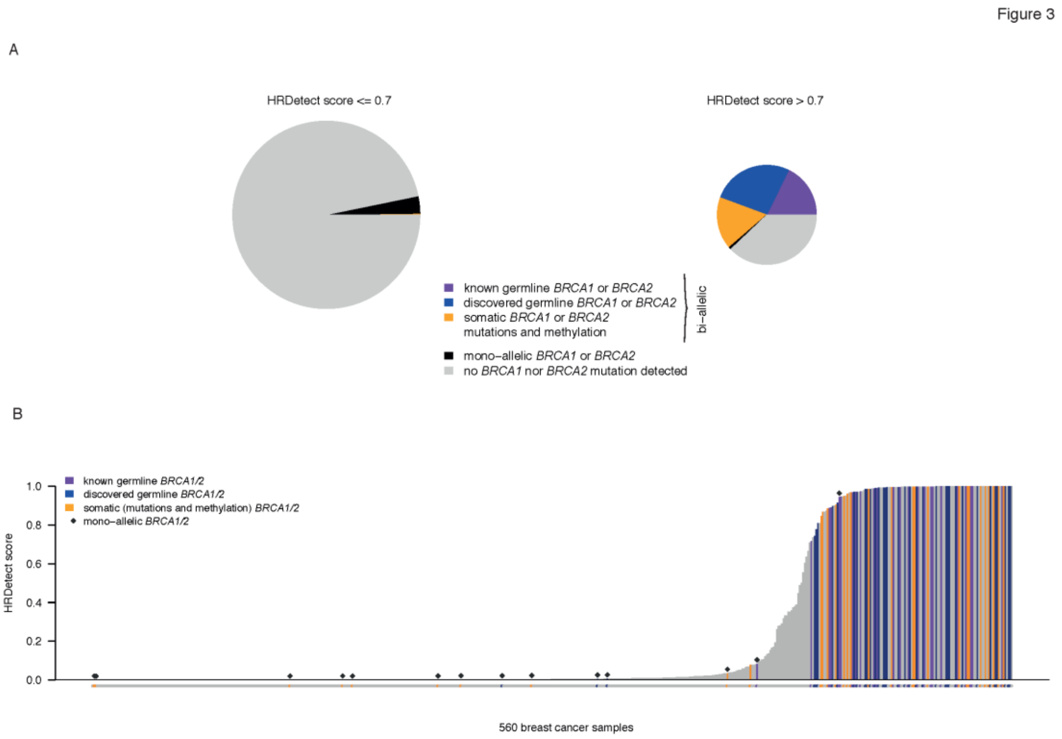


Figure 3: HRDetect as a probabilistic classifier

A. Piechart depicting the *BRCA*1 and *BRCA2* mutation status of samples in the 560 breast cancer set which produced, on the lefthand side HRDetect scores below the cut-off of 0.7 and on the right handside above 0.7. Purple = previously known germline *BRCA1* and *BRCA2* with loss of the alternative allele; blue = newly discovered gemline *BRCA1* and *BRCA2* with loss of the alternative allele; orange = somatic gemline *BRCA1* and *BRCA2* and DNA hypermethylation of *BRCA1* promoter with loss of the alternative allele; black = monoallelic germline and somatic *BRCA1* and *BRCA2* retaining the alternative allele; grey = samples in which no *BRCA1* and *BRCA2* mutation has been detected.

B. HRDetect scores for 560 breast cancer samples ordered lowest to highest scores from left to right. Coloured bars included both monoalleleic mutations and those with loss of alternative allele, purple; previously known germline *BRCA1* and *BRCA2* (24 in total of which 22 are biallelic and 2 mono-alleleic), blue; newly discovered germline *BRCA1* and *BRCA2* (36 in total of which 33 are bialleic and 3 monoalleleic), orange; somatic gemline *BRCA1* and *BRCA2* and hypermethylation of *BRCA1* promoter (31 in total of which 22 are biallelic and 9 monoalleleic), black diamonds above the bars indicate monoallelic germline and somatic *BRCA1* and *BRCA2* retaining the alternative allele (14).

Figure 4

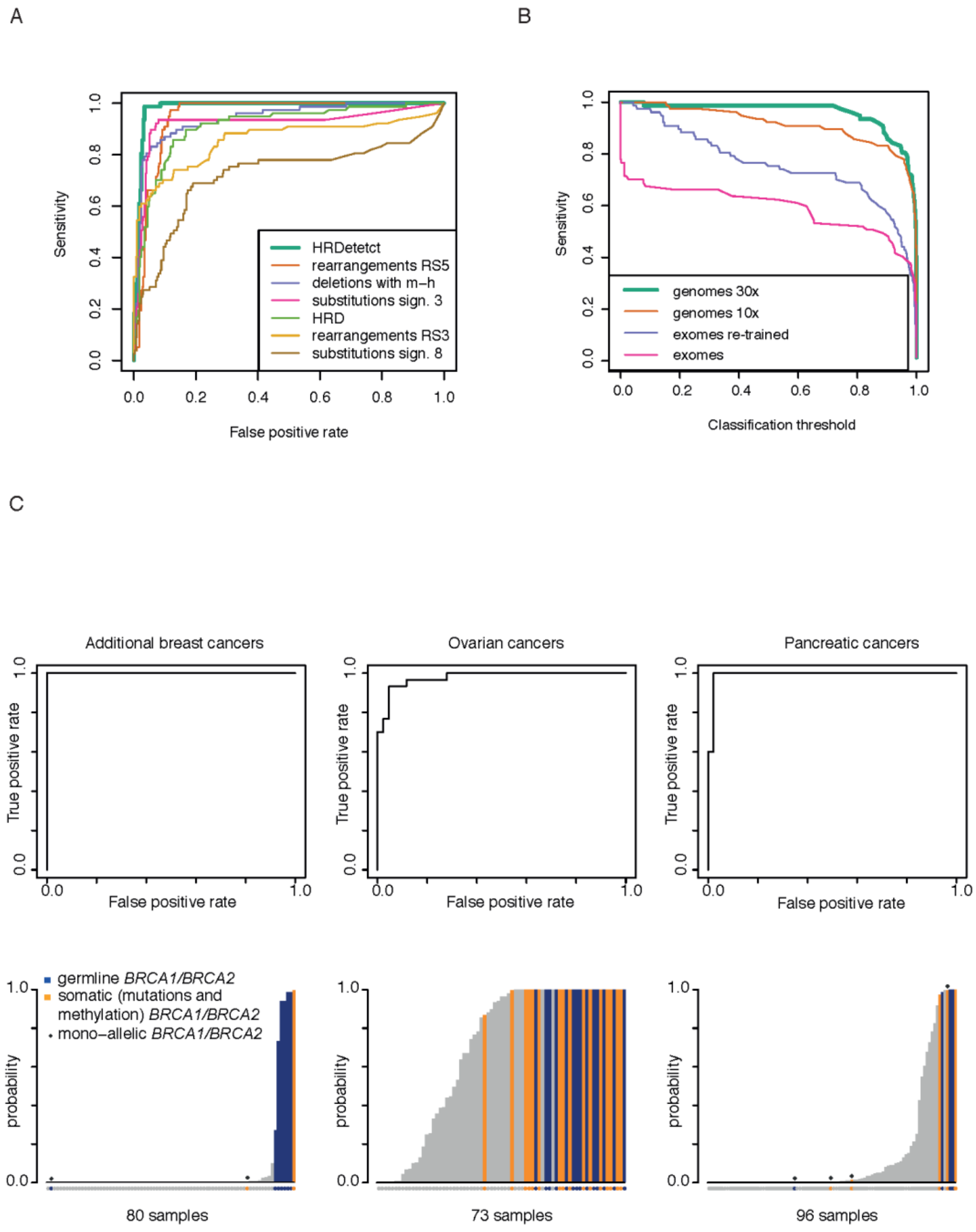


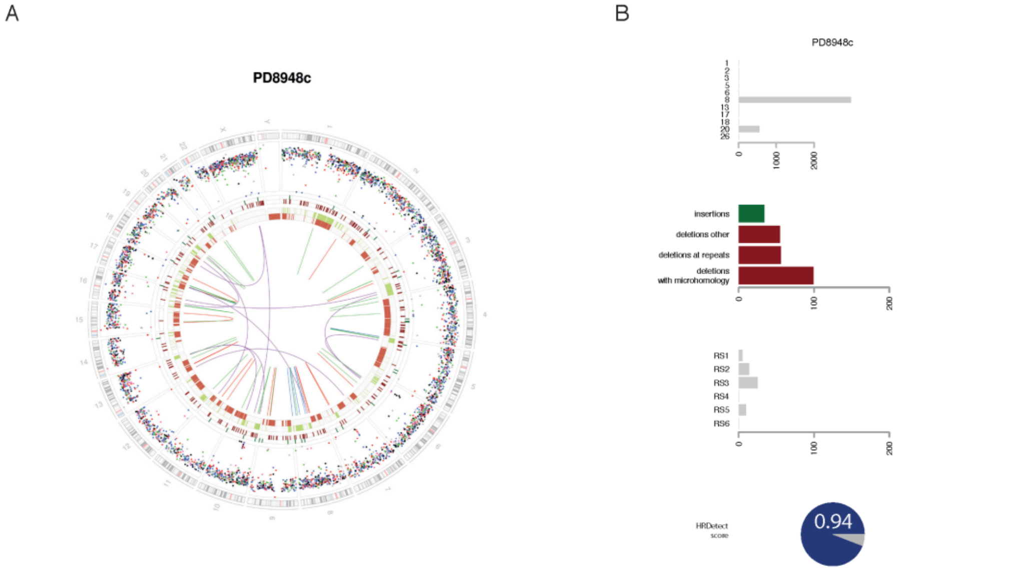
Figure 4: Performance of HRDetect and validation

A. ROC curves demonstrating the performance of HRDetect on 371 breast cancer samples as well as performance when simply using individual mutational signatures as a predictor of *BRCA1*/*BRCA2* deficiency.

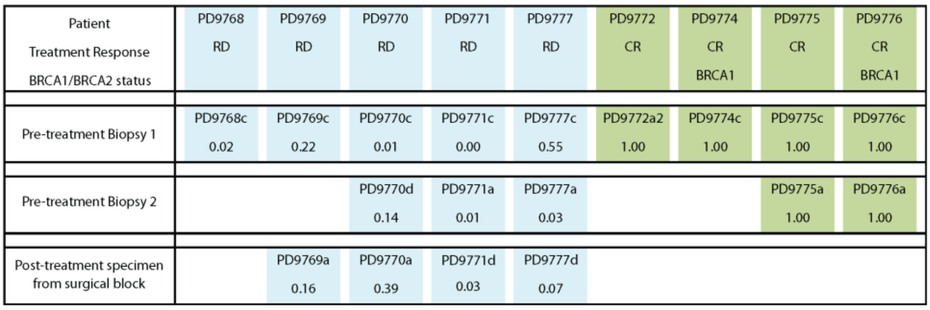
B. Comparing the sensitivity of detection of *BRCA1*/*BRCA2* deficient tumours across different types of sequencing experiments - high-coverage 30-40X genomes, low-coverage 10X genomes and whole exome sequencing (using HRDetect weights learned from WGS and retrained on WES data). 371 breast cancer samples are used in each case.

C. Performance of HRDetect on other data sets. From left to right - a cohort of 80 new breast cancers, 73 ovarian cancers and 96 pancreatic cancers. Top panel shows ROC curves for each cancer type respectively. Bottom panel shows histogram of HRDetect scores. Blue = germline *BRCA1* and *BRCA2* mutations; orange = somatic *BRCA1* and *BRCA2* mutations; grey = no *BRCA1* and *BRCA2* mutation detected; black diamonds above the bars indicate mono-allelic germline and somatic BRCA1 and BRCA2 retaining the alternative allele.

Figure 5



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
Pre-treatment 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
Pre-treatment 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				

RD = residual disease
CR = complete response
Numbers indicate HRDetect probability score for each specimen

Figure 5: Clinically relevant strengths of HRDetect

A. Genome plot from an FFPE sample from a patient with a germline *BRCA1* mutation

B. Contribution of mutation signatures, top, substitutions; middle, indels; bottom, rearrangements; and below representation of the HRDetect score.

C. HRDetect scores for nine patients treated with neoadjuvant anthracyclines +/- taxanes. Duplicate pretreatment needle biopsy samples were available for five of the samples (Pre-treatment Biospy 1 and 2). One patient (PD9770) had multifocal tumours. One patient with extremely low tumour cellularity in both biopsies and with hardly any mutations, was excluded (PD9773). HRDetect scores are provided under each sample. Blue shading indicates patients with residual disease while patients shaded green had complete response to treatment.

References

1. Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. Anglian Breast Cancer Study Group. *Br J Cancer* **83**, 1301-8 (2000).
2. John, E.M. *et al.* Prevalence of pathogenic BRCA1 mutation carriers in 5 US racial/ethnic groups. *JAMA* **298**, 2869-76 (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**, 8297-308 (2006).
4. Couch, F.J., Nathanson, K.L. & Offit, K. Two decades after BRCA: setting paradigms in personalized cancer care and prevention. *Science* **343**, 1466-70 (2014).
5. King, M.C. "The race" to clone BRCA1. *Science* **343**, 1462-5 (2014).
6. Lord, C.J. & Ashworth, A. The DNA damage response and cancer therapy. *Nature* **481**, 287-94 (2012).
7. Venkitaraman, A.R. Cancer suppression by the chromosome custodians, BRCA1 and BRCA2. *Science* **343**, 1470-5 (2014).
8. Farmer, H. *et al.* Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434**, 917-21 (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**, 913-7 (2005).
11. Fong, P.C. *et al.* Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. *N Engl J Med* **361**, 123-34 (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**, 245-51 (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**, 235-44 (2010).
14. Mateo, J. *et al.* DNA-Repair Defects and Olaparib in Metastatic Prostate Cancer. *N Engl J Med* **373**, 1697-708 (2015).
15. Ledermann, J. *et al.* Olaparib maintenance therapy in platinum-sensitive relapsed ovarian cancer. *N Engl J Med* **366**, 1382-92 (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**, 2090-8 (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**, 479-89 (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**, 1561-70 (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**, 6159-68 (2010).
25. Lord, C.J. & Ashworth, A. BRCAness revisited. *Nat Rev Cancer* **16**, 110-20 (2016).
26. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-9 (2008).
27. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* (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**, 979-93 (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**, 415-21 (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**, 489-94 (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**, 751-9 (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**, 189-96 (2010).
39. Chappuis, P.O. *et al.* A significant response to neoadjuvant chemotherapy in BRCA1/2 related breast cancer. *J Med Genet* **39**, 608-10 (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**, 4880-90 (2010).
41. Edwards, S.L. *et al.* Resistance to therapy caused by intragenic deletion in BRCA2. *Nature* **451**, 1111-5 (2008).
42. Sakai, W. *et al.* Secondary mutations as a mechanism of cisplatin resistance in BRCA2-mutated cancers. *Nature* **451**, 1116-20 (2008).

Competing Financial Interests

H. Davies, D. Glodzik and S. Nik-Zainal are inventors on a patent application encompassing the code and intellectual principle on this algorithm.

Acknowledgements

This work has been performed on data that was previously published. It was 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 the 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 N° 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).

Personal funding: DG was supported by the EU-FP7-SUPPRESSTEM project. ALR is partially supported by the Dana-Farber/Harvard Cancer Center SPORE in Breast Cancer (NIH/NCI 5 P50 CA168504-02). AS was supported by Cancer Genomics Netherlands (CGC.nl) through a grant from the Netherlands Organisation of Scientific research (NWO). CS is supported by a grant from the Breast Cancer Research Foundation. EB was funded by EMBL. GK is supported by National Research Foundation of Korea (NRF) grants funded by the Korean government (NRF 2015R1A2A1A10052578). SN-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 will not have gained these insights.

Accession codes

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

Accession numbers for the additional 80 breast cancers used for validation are:
Sequence BAM files: EGAD00001002740
SNP6 array CEL files: EGAD00010001079

Online methods

1. Dataset

Internal Review Boards of each participating institution approved collection and use of samples of all patients in this study.

DNA was extracted from 560 breast cancer cases along with corresponding normal tissue and subjected to whole genome sequencing as described previously. 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 for calling somatic substitutions. Indels in the tumour and normal genomes were called using a modified Pindel version 2.0 (http://cancerit.github.io/cgpPindel/) on the NCBI37 genome build[11]. 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[44] 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.

1.1 Mutational signatures background

Mutation signature analysis based on nonnegative matrix factorization (NMF) was performed as described previously[27,45]. Twelve base consensus 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 (characterised by C>T transitions at N$\underline{\text{C}}$G, where the underlined base is mutated) and 5 (primarily characterised by C>T and T>C mutations) have previously been associated with age[46]. Signature 2 is predominantly composed of C>T substitutions at TCN, and signature 13 predominantly comprise C>G mutations at TCN, 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 tumours. 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 aetiology of signatures 17 (characterised by T>G mutations at NTT) and 18 (exhibiting a high proportion of C>A mutations) is unknown. Signature 30 (characterised 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. RS1 and RS3 were characterised by non-clustered tandem duplications. RS1 mostly with tandem duplications of >100 kb, while, RS3 was predominantly associated with small tandem duplications <10 kb. RS2 was characterized by non-clustered deletions (>100 kb), inversions and interchromosomal translocations. RS4 was characterized by clustered interchromosomal translocations. RS5 was associated with non-clustered deletions <100 kb, while RS6 contained clustered inversions and deletions. RS5 is enriched in *BRCA1/BRCA2* null tumours and an excess of RS3 with *BRCA1* null tumours[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 >3bp in length and are characteristic of defective non-homologous end-joining based DNA double strand break repair. While 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 contribution of each mutation signature per sample.

1.2 Identifying whether a sample had particular mutational signatures

An iterative algorithm was used to identify the set of COSMIC signatures [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 neighbouring bases immediately 5' and 3′ to the mutated base and by the mutated base itself). Each mutation was orientated with respect to the pyrimidine strand and consequently each vector contained 96 elements. The algorithm started from a random exposure and iteratively moved a number of mutations from one signature to another, where both number of mutations and signatures were randomly picked. This choice was aimed to maximize the cosine similarity between observed and reconstructed vectors, and the algorithm stopped when no improvement to the cosine similarity was found in 1,000 consecutive random movements. This procedure was independently applied 100 times on each sample, and the median of these 100 exposures was considered to be the final sample exposure.

1.3 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 tumours was performed using ASCAT (v2.1.1), to generate integral allele-specific copy number profiles for the tumour cells[47]. ASCAT was also applied to next-generation sequencing data directly with highly comparable results. Resulting allele-specific data generated by ASCAT were used in the calculation of homologous recombination deficiency (HRD) index score using implementations made in R[48,49]. See Supplementary table 1 for HRD index scores.

1.4 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 Supplementary information.

2. Lasso logistic regression modelling

2.1 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 tumour sample being HR-deficient during its evolution.

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

This supervised learning method was first applied to a cohort of 22 germline mutation carriers of *BRCA1* and *BRCA2* with clear loss of the alternative parental allele. There were also 235 control samples that did not have *BRCA1*/*BRCA2* mutations or promoter hypermethylation of *BRCA1*, and any evidence of signatures of *BRCA1/BRCA2* and were believed to be sporadic breast cancers. BRCA1/BRCA2 proficient tumours 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 tumours.

Inputs into the algorithm were as follows:

- counts of mutations associated with each signature of single-base substitutions: signatures 1, 2, 3, 5, 6, 8, 13, 17, 18, 20, 26 – (Signature 30 was excluded as it involved only 1 sample)

- indels with micro-homology at indel breakpoint junction, indels at polynucleotide repeat tracts and other complex indels as proportions
- counts of rearrangements associated with each signature of rearrangements RS1-RS6
- 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 formula:

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

Equation 1

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

$$\mathrm{x}'' = \frac{x' - mean(x')}{sd(x')}$$

Equation 2

A lasso logistic regression[50] model was used to separate the two categories of patient samples: those affected or 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 genomic features most relevant to predicting *BRCA1*/*BRCA2* status through variable selection.

Coefficients β are learnt from the genomic parameters of *BRCA1*/*BRCA2*-proficient and *BRCA1*/*BRCA2*-deficient samples presented to the algorithm. Optimal coefficients are obtained by minimising the objective function[50]:

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

Equation 3

where:

$y_i$is BRCA status of a sample, $y_i = 1$for BRCA1/BRCA2 null samples, $y_i = 0$ otherwise

β_0 is the intercept, equivalent to the background log-odds of BRCAness

β is a vector of weights with one real value corresponding to each feature

p is the number of features characterising each sample

N is the number of samples

x_i^T is the vector of features characterising *i*th sample

λ is the penalty (real value) promoting the sparseness of the weights, as learnt through nested cross-validation in R package glmnet

$\|\beta\|_1$ is the L1 norm of the vector of weights, ie. the sum of absolute values of all entries of the coefficient vector

We constrained all β 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 non-negative 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 it is the dominant mutational process in the cancer or not.

Multiple mutational processes can exist in a tumour, 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 present.

Ultimately, the lasso logistic regression model is used to assign a probabilistic score to any new sample that is being analysed, using the normalised exposures of mutational processes in the sample (x_i^T) applying the parameters of the model (β) as follows:

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

Equation 4

where
C_i is the variable encoding the status of i^th sample
β_0 is the intercept weight
x_i^T is the vector encoding features of i^th sample; and
β is the vector of weights.

2.2 Robustness, stability and generalizability (on 22)

We trained the logistic regression model using a cohort of 22 germline *BRCA1*/*BRCA2* carriers with loss of the alternate allele, and a cohort of 235 sporadic tumours in this supervised analysis.

We used a ten-fold 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 for assessing 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 λ values that define the sparsity of the results.

We obtained the model coefficients across the ten folds (presented as boxplots in the Supplementary Figure 1) demonstrating that the results across the ten folds are consistently non-zero for each of the genomic parameters identified as distinguishing. The model was finally applied across all the data used in training, where the coefficients from this final run (Table 1) are also presented in red crosses in Supplementary Figure 1.

The genomic parameters and associated coefficients were identified for a λ of 0.000480 (mean 0.000891 and standard deviation 0.000803), to distinguish samples with HR deficiency in the final step:

Genomic feature	Weight
Deletions with micro-homology	5.889
HRD index	1.752
Substitutions signature 3	1.722
Rearrangements RS3	1.285
Rearrangements RS5	0.381

Table 1: Weights for the genomic features contributing to the classifier. Trained using samples from 22 know germline BRCA1/2 carriers and 235 quiescent tumours.

Finally, we assessed the stability of each coefficient through sub-sampling 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. non-zero coefficient). This was performed iteratively and out of 100 sub-sampling and training iterations, each coefficient was non-zero as shown in Table 2:

Genomic feature name	Number of times selected as non-zero coefficient (out of 100)
Deletions at micro-homology	100
HRD index	95
Substitutions signature 3	83
Rearrangements RS5	72
Substitutions signature 8	49
Rearrangements RS3	32
Deletions other	7

Table 2: Stability analysis of genomic features of BRCAness, when trained using half of data from the total of 22 know germline *BRCA1/2* carriers and 235 quiescent tumours.

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

2.3 Identifying further samples with BRCA1/BRCA1 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 steeply sigmoidal curve.

Apart from the 22 patients recruited into the study, many samples with high BRCAness scores had germline mutations that we had not known of at the time of enrolment 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 is samples with bi-allelic inactivation of *BRCA1/BRCA2* genes, whether germline or somatic, and included all such samples in a further round of training.

2.4 Retraining on 77 and defining HRDetect, a classifier of BRCA1/BRCA2 deficiency

In the final round of training, we included 77 samples with bi-allelic inactivation of *BRCA1/BRCA2*, and 234 quiescent tumours as negative examples. The number of *BRCA1/BRCA2* proficient tumours 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 robustness and generalizability of HRDetect using nested cross-validation as before (Section 2.2).

The genomic parameters and associated coefficients learnt across the 10 folds of cross-validation are shown in Supplementary Figure 3. The boxplots show variability of each coefficient, and the red crosses show values of the coefficients when training on the whole dataset.

In comparison to training with 22 known germline *BRCA* carriers only, the variability of the coefficient values across folds 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 Table 3.

Coefficient name	Number of times non-zero (out of 100)
Deletions at micro-homology	100
HRD index	92
Substitutions signature 3	99
Rearrangements RS5	67
Substitutions signature 8	81
Rearrangements RS3	61
Deletions other	15
Substitutions signature 5	13
Substitutions signature 13	6
Rearrangement signature RS1	2

Table 3: Stability analysis of genomic features of BRCAness, when trained using half of data from the total of 77 *BRCA1/2* carriers and 234 quiescent tumours.

The logistic regression model was settled on the coefficients in Table 4, with (λ) of 0.00369 (mean 0.00478 and standard deviation 0.00104):

Genomic feature	Weight
Deletions with micro-homology	2.398
Substitutions signature 3	1.611

Rearrangements RS3	1.153
Rearrangements RS5	0.847
HRD index	0.667
Substitutions signature 8	0.091

Table 4: Weights for the genomic features contributing the the classifier. Trained using samples from 77 *BRCA1/2* carriers and 234 quiescent tumours.

These parameters were finalised in our algorithm that we have called HRDetect.

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 tumours (311 samples of 560 breast cancer genomes).

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

2.5 Assessment of accuracy of the classifiers through ROC curves

For a more comprehensive assessment of 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 characterised with respect to methylation and HRD indices. We ultimately assessed the performance of HRDetect on 371 out of 560 breast cancer genomes, ignoring 2 samples with no HRD index and 187 with missing methylation data, as their BRCA 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 bi-allelic loss of *BRCA1/2*. The area under the ROC curve for the breast cancer genomes was 0.98, which we quote in the main part of the manuscript.

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 4 describes the steps involved in training, evaluation and applying HRDetect to the full data set.

2.6 Applying HRDetect to new tumour samples

Applying the predictor to a new sample requires that it has been characterised with respect to signatures of single-base substitutions, signatures of rearrangements, copy number profile and HRD score, small insertions and deletions, together with the characteristics of adjacent sequence.

Furthermore, the features of a new sample need to be normalised as in Equations 1 and 2. Table 5 contains the means and standard deviations of each feature that were taken into account based on current settings of HRDetect.

feature	mean	sd
Rearrangements RS3	1.260	1.657
Rearrangements RS5	1.935	1.483
Substitutions signature 3	2.096	3.555
Substitutions signature 8	4.390	3.179
HRD index (copy-number)	2.195	0.750
Deletions with micro-homology (proportion of all deletions)	0.218	0.090

Table 5: Mean and standard deviations of genomic features in the cohort of 560 breast cancers.

After the features of a new sample were normalised, HRDetect score is obtained by applying Equation 4, coefficients from Table 4 and intercept $\beta_0 = -3.364$.

For details of how HRDetect was applied to new breast cancer samples, down sampled genomes, breast cancer WES and WGSs from other cancer types, see Supplementary information.

43. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).

44. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-9 (2008).

45. 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 Rep* **3**, 246-59 (2013).

46. Alexandrov, L.B. *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* **47**, 1402-7 (2015).

47. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-5 (2010).

48. Abkevich, V. *et al.* Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br J Cancer* **107**, 1776-82 (2012).

49. Natrajan, R. *et al.* Characterization of the genomic features and expressed fusion genes in micropapillary carcinomas of the breast. *J Pathol* **232**, 553-65 (2014).

50. Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological* **58**, 267-288 (1996).