Markers of BRCAness in breast cancer

Weston R. Bodily\(^1\), Brian H. Shirts\(^2\), Tom Walsh\(^3,4\), Suleyman Gulsuner\(^3,4\), Mary-Claire King\(^3,4\), Alyssa Parker\(^1\), Moom Roosan\(^5\), Stephen R. Piccolo\(^1,\ast\)

\(^1\) - Department of Biology, Brigham Young University, Provo, UT, USA
\(^2\) - Department of Laboratory Medicine, University of Washington, Seattle, Washington, USA
\(^3\) - Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, Washington, USA
\(^4\) - Department of Genome Sciences, University of Washington, Seattle, Washington, USA
\(^5\) - Pharmacy Practice Department, Chapman University School of Pharmacy, Irvine, CA, USA
\(^\ast\) - Please address correspondence to S.R.P. at stephen_piccolo@byu.edu.
Abstract

Background: Mutations in BRCA1 and BRCA2 cause deficiencies in homologous recombination repair (HR), resulting in repair of DNA double-strand breaks by the alternative non-homologous end-joining pathway, which is more error prone. HR deficiency of breast tumors is important because it is associated with better response to platinum salt therapies and to PARP inhibitors. Among other consequences of HR deficiency are characteristic somatic-mutation signatures and transcriptomic patterns. The term “BRCAness” describes tumors that harbor an HR defect but have no detectable germline mutation in BRCA1 or BRCA2. A better understanding of the genes and molecular aberrations associated with BRCAness could provide mechanistic insights and guide development of targeted treatments.

Methods: Using The Cancer Genome Atlas (TCGA) genomic data from breast cancers in 1101 patients, we identified tumors with BRCAness based on somatic mutations, homozygous deletions, and hypermethylation of BRCA1 and BRCA2. We then evaluated germline mutations, somatic mutations, homozygous deletions, and hypermethylation of 24 other breast-cancer predisposition genes. Using somatic-mutation signatures, we compared these groups against tumors from 44 TCGA patients with germline mutations in BRCA1 or BRCA2. We also compared gene-expression profiles of tumors with BRCAness versus tumors from BRCA1 and BRCA2 mutation carriers. A statistical resampling approach enabled objective quantification of similarities among tumors, and dimensionality reduction enabled graphical characterizations of these relationships.

Results: Somatic-mutation signatures of tumors having a BRCA1/BRCA2 somatic mutation, homozygous deletion, or hypermethylation (n = 64) were markedly similar to each other and to tumors from BRCA1/BRCA2 germline carriers (n = 44). Furthermore, somatic-mutation signatures of tumors with germline or somatic events in BARD1 or RAD51C showed high similarity to tumors from BRCA1/BRCA2 carriers. These findings coincide with the roles of these genes in HR and support their candidacy as genes critical to BRCAness. As expected, tumors with either germline or somatic events in BRCA1 were enriched for basal gene-expression features.

Conclusions: Somatic-mutation signatures reflect the effects of HR deficiencies in breast tumors. Somatic-mutation signatures have potential as biomarkers of treatment response and to decipher the mechanisms of HR deficiency.

Keywords: Breast cancer, mutational signature, cancer subtypes, multiomic, BRCAness, expression profiles
Introduction

Approximately 1-5% of breast-cancer patients carry a pathogenic germline variant in either BRCA1 or BRCA2. These genes play important roles in homologous recombination repair (HR) of double-stranded breaks and stalled or damaged replication forks. When the BRCA1 or BRCA2 gene products are unable to perform HR, cells may resort to non-homologous end-joining, a less effective means of repairing double-stranded breaks, potentially leading to an increased rate of DNA mutations. Patients who carry biallelic loss of BRCA1 and BRCA2 due to germline variants and/or somatic mutations often respond well to poly ADP ribose polymerase (PARP) inhibitors and platinum-salt therapies, which increase the rate of DNA damage, typically causing the cells to enter programmed cell death.

The downstream effects of BRCA mutations are distinctive. For example, BRCA1 and BRCA2-mutant tumors exhibit an abundance of C-to-T transitions across the genome. Other downstream effects include characteristic transcriptional responses. For example, it has been shown that the “Basal” gene-expression subtype is enriched for tumors with BRCA1 mutations, that BRCA1 mutations are commonly found in breast tumors with triple-negative hormone-receptor status, and that gene-expression profiles may predict PARP inhibitor responses. These patterns are consistently observable, even in the presence of hundreds of other mutations in the tumors.

In 2004, Turner, et al. coined the term BRCAness to describe patients who do not have a pathogenic germline variant in BRCA1 or BRCA2 but who have developed a tumor with an impaired ability to perform HR. This category may be useful for clinical management of patients and especially for predicting treatment responses. Recent estimates suggest that the proportion of breast-cancer patients who fall into this category may be as high as 20%. Davies, et al. demonstrated an ability to categorize patients into this category with high accuracy based on high-level mutational patterns. Polak, et al. confirmed that somatic mutations, large deletions, and DNA hypermethylation of BRCA1 and BRCA2 are reliable indicators of BRCAness. They also showed a relationship between BRCAness and germline mutations in PALB2 and hypermethylation of RAD51C. However, a considerable portion of breast tumors with HR deficiency lack a known driver. Furthermore, little is known about whether the downstream effects of germline variants, somatic variants, large deletions, and hypermethylation are similar to each other or whether these effects are similar for different genes.

An underlying assumption of the BRCAness concept is that the effects of HR deficiency are similar across tumors, regardless of the genes that drive those deficiencies and despite considerable variation in genetic...
backgrounds, environmental factors, and the presence of other driver mutations. Based on this assumption—and in a quest to identify candidate markers of BRCAness—we performed a systematic evaluation of multiomic and clinical data from 1101 patients in The Cancer Genome Atlas (TCGA)\textsuperscript{24}. In performing these evaluations, we characterized each tumor using two types of molecular signature: 1) weights that represent the tumor’s somatic-mutation profile and 2) mRNA expression values for genes used to assign tumors to the PAM50 subtypes\textsuperscript{36,37}. In this way, we sought to characterize the effects of HR defects in a comprehensive yet clinically interpretable manner. To evaluate similarities among tumors based on these molecular profiles, we used a statistical-resampling approach designed to quantify similarities among patient subgroups, even when those subgroups are small, thus helping to account for rare events. We use aberration as a general term to describe germline mutations, somatic mutations, copy-number deletions, and hypermethylation events.

**Methods**

**Data preparation and filtering**

We obtained breast-cancer data from TCGA for 1101 patients in total. To determine germline-mutation status, we downloaded raw sequencing data from CGHub\textsuperscript{38} for normal (blood) samples. We limited our analysis to whole-exome sequencing samples that had been sequenced using Illumina Genome Analyzer or HiSeq equipment. Because the sequencing data files were stored in BAM format, we used Picard Tools (SamToFastq module, version 1.131, http://broadinstitute.github.io/picard) to convert the files to FASTQ format. We used the Burrows-Wheeler Alignment (BWA) tool (version 0.7.12)\textsuperscript{39} to align the sequencing reads to version 19 of the GENCODE reference genome (hg19 compatible)\textsuperscript{40}. We used sambamba (version 0.5.4)\textsuperscript{41} to sort, index, mark duplicates, and flag statistics for the aligned BAM files. In cases where multiple BAM files were available for a single patient, we used bamUtil (version 1.0.13, https://github.com/statgen/bamUtil) to merge the BAM files. When searching for relevant germline variants, we focused on 26 genes that had been included in the BROCA Cancer Risk Panel and that had a known association with breast-cancer risk (http://tests.labmed.washington.edu/BROCA)\textsuperscript{42,43}. We extracted data for these genes using bedtools (intersectBed module, version 2)\textsuperscript{44}. We used Picard Tools (CalculateHsMetrics module) to calculate alignment metrics. For exome-capture regions across all samples, the average sequencing coverage was 44.4. The average percentage of target
bases that achieved at least 30X coverage was 33.7%. The average percentage of target bases that achieved at least 100X coverage was 12.3%.

To call DNA variants, we used Freebayes (version v0.9.21-18-gc15a283) and Pindel (https://github.com/genome/pindel). We used Freebayes to identify single-nucleotide variants (SNVs) and small insertions or deletions (indels); we used Pindel to identify medium-sized insertions and deletions. Having called these variants, we used snpEff (version 4.1) to annotate the variants and GEMINI (version 0.16.3) to query the variant data. To expedite execution of these steps, we used the GNU Parallel software. The scripts and code that we used to process the germline data can be found in an open-access repository: https://bitbucket.org/srp33/tcga_germline/src.

Geneticists experienced in variant interpretation (BHS, TW, SG, MCK) further filtered the germline variants for pathogenicity using available sources of information on variants, following accepted guidelines for variant classification as previously described. Accordingly, these germline calls were independent of variant-classification calls used in prior studies of TCGA data[50; koboldtComprehensiveMolecularPortraits2012]. To assess loss of heterozygosity (LOH), we used data from Riaz, et al. They had made LOH calls for a large proportion of the breast-cancer patients in our study.

We identified somatic SNVs and indels for each patient by examining variant calls that had been made using Mutect; these variants had been made available via the Genomic Data Commons. We used the following criteria to exclude somatic variants: 1) synonymous variants 2) variants that snpEff classified as having a “LOW” or “MODIFIER” effect on protein sequence, 3) variants that SIFT and Polyphen both suggested to be benign, and 4) variants that were observed at greater than 1% frequency across all populations in ExAC. For BRCA1 and BRCA2, we examined candidate variants based on all available sources of evidence and the University of Washington, Department of Laboratory Medicine clinical database as described previously. We compared our classifications to those publicly reported in the ClinVar database when available and found complete concordance. Based on these criteria, we categorized each variant as pathogenic, likely pathogenic, variant of uncertain significance (VUS), likely benign, or benign. Then we examined the ClinVar database for evidence that VUS or likely benign variants had been classified by others as pathogenic; however, none met this criterion. To err on the side of sensitivity, we considered any BRCA1 and BRCA2 mutation to be “mutated” if it fell into our pathogenic, likely pathogenic, or VUS categories.
Using the somatic-mutation data for each patient, we derived mutation-signature profiles using the deconstructSigs (version 1.8.0) R package\(^6^1\). As input to this process, we used somatic-variant calls that had not been filtered for pathogenicity, as a way to ensure adequate representation of each signature. The output of this process was a vector for each tumor that indicated a “weight” for each signature\(^1^9\). Figures S1-S2 illustrate these weights for two tumors that we analyzed.

We downloaded DNA methylation data via the Xena Functional Genomics Explorer\(^6^2\). These data were generated using the Illumina HumanMethylation27 and HumanMethylation450 BeadChip platforms. For the HumanMethylation27 arrays, we mapped probes to genes using a file provided by the manufacturer (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL8490). For the HumanMethylation450 arrays, we mapped probes to genes using an annotation file created by Price, et al.\(^6^3\) (see http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL16304). Typically, multiple probes mapped to a given gene. Using probe-level data from BRCA1, BRCA2, PTEN, and RAD51C, we performed a preliminary analysis to determine criteria for selecting and summarizing these probe-level values. Because these genes are tumor suppressors, we started with the assumption that in most cases, the genes would be methylated at low levels. We also assumed that probes nearest the transcription start sites would be most informative. Upon plotting the data (Figure S3), we decided to limit our analysis to probes that mapped to the genome within 300 nucleotides of each gene’s transcription start site. In some cases, probes appeared to be faulty because they showed considerably different methylation levels (“beta” values) than other probes in the region (Figure S3). To mitigate the effects of these outliers, we calculated gene-level methylation values as the median beta value across any remaining probes for that gene. Then, to identify tumors that exhibited relatively high beta values—and thus could be considered to be hypermethylated—we used the getOutliersII function in the extremevalues R package (version 2.3.2)\(^6^4\) to detect outliers. When invoking this function, we specified the following non-default parameter values: distribution = "exponential", alpha = c(0.000001, 0.000001).

We downloaded copy-number-variation data from the Xena Functional Genomics Explorer\(^6^2\). These data had been generated using Affymetrix SNP 6.0 arrays; CNV calls had been made using the GISTIC2 method\(^6^5\). The CNV calls had also been summarized to gene-level values using integer-based discretization. We focused on tumors with a gene count of “−2”, which indicates a homozygous deletion.

We used RNA-Sequencing data that had been aligned and summarized to gene-level values using the original TCGA pipeline\(^2^4\). To facilitate biological and clinical interpretation, we limited the gene-expression data to The Prosigna™ Breast Cancer Prognostic Gene Signature (PAM50) genes\(^6^6\). Netanely, et al. had previously
published PAM50 subtypes for TCGA breast cancer samples; we reused this information in our study\textsuperscript{67}. For each of these genes, we also sought to identify tumors with unusually low expression levels. To do this, we used the `getOutliersI` function in the `extremevalues` package to identify outliers. We used the following non-default parameter values: \( \text{alpha} = c(0.000001, 0.000001) \), \( \text{distribution} = "\text{lognormal}" \), \( \text{FLim} = c(0.1, 0.9) \).

We parsed demographic, histopathological, and surgical variables for TCGA samples from the repository prepared by Rahman, et al.\textsuperscript{68}. We obtained drug-response data from the TCGA legacy archive (https://portal.gdc.cancer.gov/legacy-archive) and standardized drug names using synonyms from the National Cancer Institute Thesaurus\textsuperscript{69}.

\textbf{Quantitative analysis and visualization}

To prepare, analyze, and visualize the data, we wrote computer scripts in the R programming language\textsuperscript{70}. In writing these scripts, we used the following packages: `readr`\textsuperscript{71}, `dplyr`\textsuperscript{72}, `ggplot2`\textsuperscript{73}, `tidyr`\textsuperscript{74}, `reshape2`\textsuperscript{75}, `ggrepel`\textsuperscript{76}, `cowplot`\textsuperscript{77}, `data.table`\textsuperscript{78}, `UpSetR`\textsuperscript{79}, `BSgenome.Hsapiens.UCSC.hg38`\textsuperscript{80,81}, and `Rtsne`\textsuperscript{82}. We created a series of R scripts that execute all steps of our analysis and generate the figures in this paper; these documents are available at https://osf.io/9jhr2.

To reduce data dimensionality for visualization purposes, we applied the Barnes-Hut t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm\textsuperscript{83,84} to the mutation signatures and PAM50 expression profiles. This reduced the data to two dimensions, which we plotted as Cartesian coordinates. To quantify homogeneity within a group of tumors that harbored a particular aberration, we calculated the pairwise Euclidean distance between each patient pair in the group and then calculated the median pairwise distance\textsuperscript{85}. When comparing two groups, we used a similar approach but instead calculated the median distance between each pair of individuals in either group. To determine whether the similarity within or between groups was statistically significant, we used a permutation approach. We randomized the patient identifiers, calculated the median pairwise distance within (or between) groups, and repeated these steps 10,000 times. This process resulted in an empirical null distribution against which we compared the actual median distance. We then derived empirical p-values by calculating the proportion of randomized median distances that were larger than the actual median distance.
Results

We used clinical and molecular data from breast-cancer patients in TCGA to evaluate the downstream effects of BRCA1 and BRCA2 germline mutations. We evaluated two types of downstream effect: 1) expression levels of genes that are used to classify tumors into the PAM50 subtypes\(^ {36,37}\) and 2) signatures that reflect a tumor’s overall somatic-mutation profile in a trinucleotide context\(^ {18,19}\). We used expression data for the PAM50 genes due to their biological and clinical relevance. We used somatic-mutation signatures because they reflect the genomic effects of HR defects and have been associated with BRCA1/BRCA2 mutation status\(^ {18,19}\). First, we assessed whether either of these profile types are more homogeneous in BRCA1/BRCA2 germline carriers than in randomly selected patients. Next we evaluated the robustness of potential criteria for classifying tumors into the “BRCAAness” category. These criteria included somatic mutations, homozygous deletions, and DNA hypermethylation of BRCA1 and BRCA2. Similarly, we assessed whether these types of aberration in 24 other breast-cancer predisposition genes have similar effects to BRCA1/BRCA2 aberrations. Before classifying any gene as a candidate BRCAAness gene, we required that the effects of these aberrations be consistent across multiple aberration types.

Of 993 breast-cancer patients with available germline data, 22 harbored a pathogenic SNV or indel in BRCA1; 22 harbored a BRCA2 variant (Figure 1A). All but 3 BRCA1 carriers and all but 7 BRCA2 carriers experienced loss of heterozygosity (LOH) in the same gene (Figures S4-S5). BRCA1 carriers fell into the “Basal” (n = 17); Her2 (n = 1), Luminal A (n = 2), and Luminal B (n = 1) gene-expression subtypes (Figure S6)\(^ {21,36,37}\). Most BRCA2 carriers fell into the Luminal A subtype (n = 13); the remaining individuals were dispersed across the other subtypes. As demonstrated previously\(^ {19}\), the primary somatic-mutation signature for most BRCA1 and BRCA2 carriers was “Signature 3”; however, other signatures (especially 1A) were also common (Figure S7). Figure S8 shows the overlap between these two types of molecular profile.

Although it is useful to evaluate breast-cancer patients based on the primary subtype or signature associated with each tumor, tumors are aggregates of multiple subtypes and signatures. To account for this diversity, we characterized tumors based on 1) gene-expression levels for all available PAM50 genes and 2) all 27 somatic-mutation signatures. To enable visualization of these profiles, we used the t-SNE technique to reduce the dimensionality of these profiles. Generally, tumors with the same primary subtype or signature clustered together in these visualizations (Figures 2-3); however, in some cases, this did not happen. For example, the dimensionally reduced gene-expression profiles for Basal tumors formed a tight, distinct cluster (Figure ??). But some Basal tumors were distant from this cluster, and one “Normal-like” tumor was located in this
cluster. Similarly, tumors assigned to somatic-mutation “Signature 3” formed a cohesive cluster (Figure 3), but some “Signature 3” tumors were separate. These observations highlight the importance of evaluating molecular profiles as a whole, not just using a single, primary category.

Under the assumption that BRCA1/BRCA2 germline variants exhibit recognizable effects on tumor transcription, we used a statistical-resampling approach (see Methods) to evaluate whether tumors from BRCA1 carriers have homogeneous gene-expression profiles. As expected based on the tumors’ primary PAM50 classifications, 18 of 22 BRCA1 carriers overlapped closely with the Basal subtype (Figure 4A). But as a whole, the expression profiles for this group were not more homogeneous than expected by random chance (p = 0.065; Figure S9A), perhaps because the 4 non-Basal samples exhibited gene-expression profiles that were vastly different from the Basal tumors. Similarly, BRCA2 carriers were not significantly homogeneous (p = 0.16; Figure S9B); tumors from these individuals were dispersed across the gene-expression topography (Figure 4B). In contrast, somatic-mutation signatures of BRCA1 germline carriers were more homogeneous than expected by chance (p = 0.0004; Figures 5A and S10A), as were those from BRCA2 carriers (p = 0.0034; Figures 5B and S10B). None of the three BRCA1 carriers who lacked LOH events clustered closely with the remaining BRCA1 tumors (Figure 5A). Of the 7 BRCA2 tumors without detected LOH events, 4 were among those that failed to cluster closely with the remaining BRCA2 tumors (Figure 5B). These observations confirm that germline BRCA1/BRCA2 mutations leave a recognizable imprint on a tumor’s mutational landscape but that this imprint is more likely in combination with a second “hit” in the same gene.19,32,86

Next we evaluated similarities between BRCA1 and BRCA2 germline carriers. Although some BRCA2 carriers fell into the Basal gene-expression subtype, overall profiles for these patients were dissimilar to those from BRCA1 carriers (p = 0.99; Figures 4A-B and S11A). However, the opposite held true for somatic-mutation signatures: tumors from BRCA1 and BRCA2 carriers were highly similar to each other (p = 0.0001; Figures 5A-B and S12A).

A somatic mutation, homozygous deletion, or DNA hypermethylation occurred in BRCA1 and BRCA2 for 64 patients (Figure 1B-D). Most of these events were mutually exclusive with each other and with germline variants (Figure S13). Whether for PAM50 subtypes or somatic-mutation signatures, tumors with BRCA1 hypermethylation were relatively homogeneous and highly similar to tumors from BRCA1 germline carriers (Figures 4G, 5G, S9G, S10G; Table 1). For PAM50 gene expression, no other aberration type showed significant similarity to BRCA1 germline mutations. Somatic-mutation signatures from tumors with BRCA1 somatic mutations or homozygous deletions were significantly similar to those from BRCA1 germline
mutations (Table 1). Only 2 tumors had BRCA2 hypermethylation, but the mutational signatures for these samples were significantly similar to tumors from BRCA2 germline carriers (p = 0.0014; Figure 5H).

Likewise, BRCA2 somatic mutations and homozygous deletions produced mutational signatures that were similar to germline BRCA2 carriers (Table 1; Figures 5D and 5F). Based on these findings, we conclude that disruptions of BRCA1 and BRCA2 exert similar effects on somatic-mutation signatures—but not PAM50 gene expression—whether those disruptions originate in the germline or via somatic processes. To provide further evidence, we aggregated all patients who had any type of BRCA1 or BRCA2 aberration into a BRCAness reference group. As a whole, mutational signatures for this group were much more homogeneous than expected by chance (p = 0.0001; Figure S14). We used this reference group to evaluate other criteria that might classify patients into the BRCAness category. For our remaining evaluations, we used somatic-mutation signatures—rather than PAM50 gene expression—for these assessments because they coincided so consistently with BRCA aberration status, in line with the definition of BRCAness as an HR defect.

We examined data for 24 additional breast-cancer predisposition genes and evaluated whether molecular aberrations in these genes result in mutational signatures that are similar to our BRCAness reference group. We found pathogenic and likely pathogenic germline mutations in 15 genes. The most frequently mutated were CHEK2, ATM, and NBN (Figures S15 and S16). We found potentially pathogenic somatic mutations in all 24 genes, most frequently in TP53, CDH1, and PTEN (Figures S17 and S18). Homozygous deletions occurred most frequently in PTEN, CDH1, and CHEK1 (Figures S19 and S20). Finally, 5 genes were hypermethylated (Figures S21 and S22). Typically, these events were rare for a given gene. Using our resampling approach, we compared each aberration type in each gene against the BRCAness reference group. In cases where an aberration overlapped between the reference and comparison groups, we removed individuals who harbored that aberration. For 8 genes, at least one type of aberration attained statistical significance (Table 2). A total of 8 aberrations occurred in BARD1 across 3 categories of aberration; all 3 categories were statistically significant (Table 2). RAD51C homozygous deletions (n = 2) and hypermethylation (n = 32) attained significance, but germline mutations (n = 1) and somatic mutations (n = 3) did not. TP53 homozygous deletions (n = 15) were significant, but somatic mutations (n = 302) and germline mutations (n = 2) were not.

Lastly, we evaluated the following types of data for candidacy as BRCAness markers: 1) unusually low mRNA expression, 2) demographic, histopathological, and surgical observations, and 3) patient drug responses. First, we calculated the median Euclidean distance—based on somatic-mutation
signatures—between each patient and the BRCAness reference group. Then we used a two-sided Pearson correlation test to assess the relationship between these median distances and each candidate variable. In determining whether a tumor exhibited unusually low mRNA expression for a given gene, we used an outlier-detection technique (see Methods). Unusually low expression of RAD51C (rho = 0.29, p = 4.9e-6) and BRCA1 (rho = 0.26, p = 4.2e-5) showed the strongest positive correlation with the reference group, whereas BARD1 (rho = -0.28, p = 8.5e-5) and CDH1 (rho = -0.28, p = 8.5e-5) showed the strongest negative correlation (Figures S23 and 6). Triple-negative status, infiltrating ductal carcinoma histology, and close surgical margins were the most positively associated clinical variables (Figure S24). No chemotherapy treatment was significantly associated with BRCAness, though sample size (n = 211) was relatively small for the drug data (Figure S25).

Discussion

By definition, BRCAness tumors have HR defects. As with germline mutations in BRCA1 and BRCA2, these deficiencies could be exploited therapeutically. Various criteria have been proposed as indicators of BRCAness, including triple-negative hormone-receptor status, somatic mutations in BRCA1, hypermethylation of BRCA1, germline mutations in PALB2, and hypermethylation of RAD51C. However, relatively little has been understood about whether these aberrations are reliable indicators of BRCAness, whether these aberrations have similar downstream effects as germline BRCA1/BRCA2 mutations, or whether aberrations in other genes in the HR pathway could be used as reliable markers of BRCAness. We evaluated these questions using a publicly available, multiomic dataset and used robust, quantitative methods to evaluate the downstream effects of these aberrations. Our permutation approach takes multiple variables (e.g. the full profile of signature weights) into account simultaneously, not just the primary subtype. Although we observed a clear relationship between germline BRCA1 mutations and the “Basal” gene-expression subtype—which overlaps considerably with triple-negative status—we otherwise observed few consistent patterns in the gene-expression data. In contrast, we observed clear and consistent patterns for the somatic-mutation signatures. Thus we conclude that somatic-mutation signatures are more useful indicators of BRCAness than gene-expression levels. Germline BRCA1 mutations affected somatic-mutation signatures similarly to germline BRCA2 mutations. Furthermore, somatic-mutations, homozygous deletions, and hypermethylation of BRCA1 and BRCA2 had downstream effects similar to germline mutations in these genes. As a whole, tumors with any
BRCA1/BRCA2 aberration formed a cohesive group, against which we compared other tumors. For a gene to be considered a strong BRCAness biomarker candidate, we required that at least two types of molecular aberration show significant similarity to the BRCAness reference group, suggesting that aberrations in the gene leave a recognizable imprint on the somatic-mutation landscape. This allowed us to derive insights even though a single type of aberration may have occurred rarely in a given gene. Two genes met these criteria: BARD1 and RAD51C. These genes both form a complex with BRCA1 to help repair double-stranded breaks via homologous recombination; both proteins are enriched in triple-negative breast tumors. Our findings provide additional evidence that defects in these genes have interchangeable effects on HR and that the functional status of these genes are a reliable indicator of BRCAness. BRCA2 interacts with RAD51 as well as PALB2.

Some genes showed significant similarity to the BRCAness reference group for one type of aberration only (Table 2). These included germline mutations in PALB2 and RAD51B, which have a clear mechanistic link to BRCA1 and BRCA2. Determining which germline mutations are pathogenic remains a challenging task, so it is possible that more- or less-stringent filtering of candidate aberrations would lead to more consistent results. In addition, it is likely that mono-allelic inactivation of these and other genes may be insufficient to impair HR function. Tumors with homozygous deletions in TP53 were significantly similar to the BRCAness groups; somatic mutations in this gene showed considerable overlap with the BRCAness tumors, but this similarity did not reach statistical significance. TP53 has long been recognized as an important gene in breast cancer, and mutations in this gene have been shown to associate with germline mutations in BRCA1 and BRCA2. However, because TP53 mutations occur frequently in breast cancer overall, they may be sensitive but non-specific biomarkers of BRCAness. Perhaps TP53 aberrations act as secondary events that compromise genomic integrity in combination with initiating events in the HR pathway.

Although the mutational-signature patterns we observed were highly consistent in many cases, it remains to be determined whether these observations are clinically relevant. Clinical trials are currently underway to identify biomarkers for carboplatin, a platinum-salt agent. Tutt, et al. concluded that BRCA1/BRCA2 mutations and triple-negative hormone status were reliable biomarkers of objective treatment responses but that BRCA1 hypermethylation was not. It may be that other BRCAness genes or different types of aberration will become useful markers of treatment response.

Our statistical-resampling approach uses Euclidean distances to evaluate similarity (see Methods). For visualization, we used a two-dimensional representation of the same data. In most cases, these two methods...
led to similar conclusions. However, we placed most confidence in the empirical p-values calculated using our resampling approach, even if those conclusions differed from what we observed visually.

Conclusions

Altogether our findings shed new light on factors that may be useful to classify patients into the BRCAness category and demonstrate an objective methodology for categorizing tumor subtypes, in general.
Figure 1: Molecular aberrations in BRCA1 and BRCA2 across all breast-cancer patients. A) Germline mutations, B) Somatic mutations, C) copy-number variations, D) DNA methylation levels. SNV = single nucleotide variation.
Figure 2: Two-dimensional representation of PAM50 gene-expression levels. We obtained expression levels for the PAM50 genes and used the t-distributed Stochastic Neighbor Embedding (t-SNE) method to reduce the data to two dimensions. Each point on the plot represents a single tumor, overlaid with colors that represent the tumor’s primary PAM50 subtype. Generally, the PAM50 subtypes clustered cohesively, but there were exceptions. For example, some
Basal tumors (A) exhibited expression patterns that differed considerably from the remaining Basal tumors. The normal-like tumors (E) showed the most variability in expression.
Figure 3: Two-dimensional representation of somatic-mutation signatures. We summarized each tumor based on their somatic-mutation signatures, which represent overall mutational patterns in a trinucleotide context. We used the t-distributed Stochastic Neighbor Embedding (t-SNE) method to reduce the data to two dimensions. Each point on the plot represents a single tumor, overlaid with colors that represent the tumor’s primary somatic-mutation signature. Mutational Signature 1A (A) was the most prevalent; these tumors were widely dispersed across the signature.
Signatures 1B (B), 2 (C), and 3 (D) were relatively small and formed cohesive clusters. The remaining 23 clusters were rare individually and were dispersed broadly.
Figure 4: *BRCA1* and *BRCA2* aberrations on the PAM50 gene-expression landscape. Using the same two-dimensional representation of PAM50 gene-expression levels shown in Figure 2, this plot indicates which patients had germline mutations (A, B), somatic mutations (C, D), homozygous deletions (E, F), or hypermethylation events (G, H) in *BRCA1* and *BRCA2*, respectively. Many of these tumors overlapped with the Basal subtype, but other tumors
were dispersed broadly across the gene-expression landscape. Diamonds represent tumors with multiple aberrations of a given type.
Figure 5: *BRCA1* and *BRCA2* aberrations on the somatic-mutation signature landscape. Using the same two-dimensional representation of mutational signatures shown in Figure 3, this plot indicates which patients had germline mutations (A, B), somatic mutations (C, D), homozygous deletions (E, F), or hypermethylation events (G, H) in *BRCA1* and *BRCA2*, respectively. Largely, these tumors had similar somatic-mutation signatures. Diamonds represent tumors with multiple aberrations of a given type.
Low mRNA expression

Patient
Low mRNA expression

Figure 6: Relationship between BRCA aberration status and relatively low gene expression. We identified tumors with relatively low expression for cancer-predisposition genes (see Figure S23) and evaluated whether the somatic-mutation signatures of these tumors were relatively similar or dissimilar to tumors with a BRCA aberration. Low expression of RAD51C and BRCA1 showed the strongest positive correlation between gene-expression status and the BRCAness reference group. Low expression of BARD1 and CDH1 showed the strongest negative correlation between gene-expression status and the BRCAness reference group.
Table 1: Results of similarity comparisons among BRCA aberration groups. We compared PAM50 gene-expression levels or somatic-mutation signatures between groups of patients who harbored aberrations in *BRCA1* or *BRCA2*. We evaluated whether patients in one group (e.g., those who harbored a *BRCA1* germline mutation) were more similar to patients in a second group (e.g., those with *BRCA2* germline mutation) than random patient subsets of the same sizes. The numbers in this table represent empirical p-values. In cases where an individual harbored an aberration in both comparison groups, we excluded that patient from the comparison.

| Aberration Type 1                  | Aberration Type 2                  | PAM50 Subtypes | Mutational Signatures |
|-----------------------------------|-----------------------------------|----------------|-----------------------|
| BRCA1 germline mutation (n = 22)  | BRCA2 germline mutation (n = 22)  | 0.997          | 1e-04                 |
| BRCA1 germline mutation (n = 22)  | BRCA1 somatic mutation (n = 14)   | 0.1203         | 1e-04                 |
| BRCA1 germline mutation (n = 22)  | BRCA1 homozygous deletion (n = 8) | 0.924          | 0.0246                |
| BRCA1 germline mutation (n = 22)  | BRCA1 hypermethylation (n = 16)   | 0.0182         | 1e-04                 |
| BRCA2 germline mutation (n = 22)  | BRCA2 somatic mutation (n = 12)   | 0.8818         | 0.0013                |
| BRCA2 germline mutation (n = 22)  | BRCA2 homozygous deletion (n = 19)| 0.6394         | 1e-04                 |
| BRCA2 germline mutation (n = 22)  | BRCA2 hypermethylation (n = 2)    | 0.6855         | 0.0014                |
Table 2: Summary of comparisons between the BRCA1/11 reference group and groups of patients who harbored a specific type of aberration in a candidate BRCA1/11 gene. We evaluated whether somatic-mutation signatures from patients who harbored a given type of aberration (e.g., BARD1 germline mutation) were more similar to the BRCA1/11 reference group than expected by random chance. The numbers in this table represent empirical p-values. In cases where no patient had a given type of aberration in a given gene, we list “N/A”. The “Any” group represents individuals who harbored any type of aberration in a given gene.

| Gene     | Germline mutation | Somatic mutation | Homozygous deletion | Hypermethylation | Any             |
|----------|-------------------|------------------|---------------------|------------------|-----------------|
| BARD1    | 1e-04 (n = 1)     | 1e-04 (n = 2)    | 4e-04 (n = 5)       | N/A              | 1e-04 (n = 8)   |
| CTNNA1   | N/A               | 0.991 (n = 8)    | 2e-04 (n = 6)       | N/A              | 0.6149 (n = 14) |
| FAM175A  | N/A               | 0.993 (n = 2)    | 2e-04 (n = 3)       | N/A              | 0.2417 (n = 5)  |
| PALB2    | 0.0098 (n = 3)    | 0.8695 (n = 5)   | N/A                 | N/A              | 0.3641 (n = 8)  |
| PTEN     | 0.9594 (n = 1)    | 0.9986 (n = 51)  | 0.0203 (n = 56)     | 0.7675 (n = 2)   | 0.797 (n = 110) |
| RAD51B   | 0.0013 (n = 3)    | 0.5743 (n = 3)   | 0.3831 (n = 9)      | N/A              | 0.2595 (n = 15) |
| RAD51C   | 0.0469 (n = 1)    | 0.9848 (n = 3)   | 0.0151 (n = 2)      | 0.0012 (n = 32)  | 0.0027 (n = 38) |
| TP53     | 0.9246 (n = 2)    | 0.0747 (n = 302) | 0.0015 (n = 15)     | N/A              | 0.0751 (n = 319) |
Declarations

Ethics approval and consent to participate

Brigham Young University’s Institutional Review Board approved this study under exemption status. This study uses data collected from public repositories only. We played no part in patient recruiting or in obtaining consent. We have adhered to guidelines from TCGA on handling data.

Consent for publication

Not applicable.

Availability of data and material

The datasets generated and analyzed during the current study are available in the Open Science Framework repository (https://osf.io/9jhr2). (We are not permitted to share the germline-mutation data.)

Competing interests

TW consults for Color Genomics. Otherwise, the authors declare that they have no competing interests.

Funding

Funding for this study was provided through Brigham Young University Graduate Studies and the Simmons Center for Cancer Research. In addition, we acknowledge grant support from NIH 1R35CA197458, Komen Foundation SAC110020, and Breast Cancer Research Foundation BCRF18-088.

Author’s contributions

WRB and SRP conceived the study design, prepared and analyzed data, and interpreted results. BHS, TW, SG, and MCK evaluated variant pathogenicity and contributed intellectual insights regarding study design and data interpretation. AP and MR parsed and evaluated the pharmacological data. WRB and SRP wrote the manuscript. BHS, TW, MCK, AP and MR edited the manuscript.
Acknowledgements

Results from this study are in part based upon data generated by TCGA and managed by the United States National Cancer Institute and National Human Genome Research Institute (see http://cancergenome.nih.gov).

We thank the patients who participated in this study and shared their data publicly. We thank the Fulton Supercomputing Laboratory at Brigham Young University for providing computational facilities.
References

1. Hall, J. M. *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science (New York, N.Y.)* **250**, 1684–9 (1990).

2. Moynahan, M. E., Chiu, J. W., Koller, B. H. & Jasin, M. Brca1 Controls Homology-Directed DNA Repair. *Molecular Cell* **4**, 511–518 (1999).

3. Stratton, M. R. & Rahman, N. The emerging landscape of breast cancer susceptibility. *Nature genetics* **40**, 17–22 (2008).

4. John, E. M. *et al.* Prevalence of pathogenic BRCA1 mutation carriers in 5 US racial/ethnic groups. *JAMA* **298**, 2869–2876 (2007).

5. 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 Research* **66**, 8297–8308 (2006).

6. Li, X. & Heyer, W.-D. Homologous recombination in DNA repair and DNA damage tolerance. *Cell Research* **18**, 99–113 (2008).

7. O’Donovan, P. J. & Livingston, D. M. BRCA1 and BRCA2: Breast/ovarian cancer susceptibility gene products and participants in DNA double-strand break repair. *Carcinogenesis* **31**, 961–7 (2010).

8. Lord, C. J. & Ashworth, A. The DNA damage response and cancer therapy. *Nature* **481**, 287–294 (2012).

9. Tutt, A. *et al.* Mutation in Brca2 stimulates error-prone homology-directed repair of DNA double-strand breaks occurring between repeated sequences. *The EMBO Journal* **20**, 4704–4716 (2001).

10. Xia, F. *et al.* Deficiency of human BRCA2 leads to impaired homologous recombination but maintains normal nonhomologous end joining. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 8644–8649 (2001).

11. Moynahan, M. E., Pierce, A. J. & Jasin, M. BRCA2 is required for homology-directed repair of chromosomal breaks. *Molecular Cell* **7**, 263–272 (2001).

12. Moore, K. *et al.* Maintenance Olaparib in Patients with Newly Diagnosed Advanced Ovarian Cancer. *The New England Journal of Medicine* **379**, 2495–2505 (2018).
13. Robson, M. et al. Olaparib for Metastatic Breast Cancer in Patients with a Germline BRCA Mutation. *The New England Journal of Medicine* **377**, 523–533 (2017).

14. Litton, J. K. et al. Talazoparib in Patients with Advanced Breast Cancer and a Germline BRCA Mutation. *The New England Journal of Medicine* **379**, 753–763 (2018).

15. Commissioner, O. of the. Press Announcements - FDA approves first treatment for breast cancer with a certain inherited genetic mutation.

16. 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 (London, England)* **376**, 235–244 (2010).

17. Fong, P. C. et al. Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. *The New England Journal of Medicine* **361**, 123–134 (2009).

18. Nik-Zainal, S. et al. Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* 979–993 (2012). doi:10.1016/j.cell.2012.04.024

19. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–21 (2013).

20. Zou, X. et al. Validating the concept of mutational signatures with isogenic cell models. *Nature Communications* **9**, 1744 (2018).

21. Sorlie, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98**, 10869–10874 (2001).

22. Foulkes, W. D. et al. Germline BRCA1 mutations and a basal epithelial phenotype in breast cancer. *Journal of the National Cancer Institute* **95**, 1482–1485 (2003).

23. Lee, E. et al. Characteristics of triple-negative breast cancer in patients with a BRCA1 mutation: Results from a population-based study of young women. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **29**, 4373–4380 (2011).

24. Koboldt, D. C. et al. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).

25. Sørlie, T. et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* **100**, 8418–8423 (2003).
26. Foulkes, W. D. et al. The prognostic implication of the basal-like (cyclin E high/p27 low/p53+/glomeruloid-microvascular-proliferation+) phenotype of BRCA1-related breast cancer. *Cancer Research* **64**, 830–835 (2004).

27. Severson, T. M. et al. The BRCA1ness signature is associated significantly with response to PARP inhibitor treatment versus control in the I-SPY 2 randomized neoadjuvant setting. *Breast Cancer Research* **19**, 99 (2017).

28. Wood, L. D. et al. The genomic landscapes of human breast and colorectal cancers. *Science (New York, N.Y.)* **318**, 1108–13 (2007).

29. Turner, N., Tutt, A. & Ashworth, A. Hallmarks of 'BRCAness' in sporadic cancers. *Nature reviews. Cancer* **4**, 814–9 (2004).

30. Lord, C. J. & Ashworth, A. BRCAness revisited. *Nature Reviews Cancer* **16**, 110–120 (2016).

31. Davies, H. et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nature Medicine* **23**, 517–525 (2017).

32. Polak, P. et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nature Genetics* **49**, 1476–1486 (2017).

33. Woodward, A. M., Davis, T. A., Silva, A. G. S., Kirk, J. A. & Leary, J. A. Large genomic rearrangements of both BRCA2 and BRCA1 are a feature of the inherited breast/ovarian cancer phenotype in selected families. *Journal of Medical Genetics* **42**, e31–e31 (2005).

34. Hogervorst, F. B. L. et al. Large Genomic Deletions and Duplications in the BRCA1 Gene Identified by a Novel Quantitative Method. *Cancer Research* **63**, 1449–1453 (2003).

35. Esteller, M. et al. Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. *Journal of the National Cancer Institute* **92**, 564–9 (2000).

36. Perou, C. M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747–52 (2000).

37. Parker, J. S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **27**, 1160–7 (2009).

38. Wilks, C. et al. The Cancer Genomics Hub (CGHub): Overcoming cancer through the power of torrential data. *Database: The Journal of Biological Databases and Curation* **2014**, (2014).
39. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754–60 (2009).

40. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research* **22**, 1760–1774 (2012).

41. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: Fast processing of NGS alignment formats. *Bioinformatics (Oxford, England)* **31**, 2032–2034 (2015).

42. Walsh, T. *et al.* Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 12629–12633 (2010).

43. Shirts, B. H. *et al.* Improving performance of multigene panels for genomic analysis of cancer predisposition. *Genetics in Medicine: Official Journal of the American College of Medical Genetics* **18**, 974–981 (2016).

44. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**, 841–842 (2010).

45. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907 [q-bio]* (2012).

46. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).

47. Paila, U., Chapman, B. A., Kirchner, R. & Quinlan, A. R. GEMINI: Integrative exploration of genetic variation and genome annotations. *PLoS computational biology* **9**, e1003153 (2013).

48. Tange, O. GNU Parallel - The Command-Line Power Tool.;login: The USENIX Magazine **36**, 42–47 (2011).

49. Shirts, B. H. *et al.* Improving performance of multigene panels for genomic analysis of cancer predisposition. *Genetics in Medicine: Official Journal of the American College of Medical Genetics* **18**, 974–981 (2016).

50. Lu, C. *et al.* Patterns and functional implications of rare germline variants across 12 cancer types. *Nature Communications* **6**, 10086 (2015).
51. Riaz, N. et al. Pan-cancer analysis of bi-allelic alterations in homologous recombination DNA repair genes. *Nature Communications* **8**, 857 (2017).

52. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* **31**, 213–219 (2013).

53. Jensen, M. A., Ferretti, V., Grossman, R. L. & Staudt, L. M. The NCI Genomic Data Commons as an engine for precision medicine. *Blood* blood–2017–03–735654 (2017). doi:10.1182/blood-2017-03-735654

54. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* **4**, 1073–81 (2009).

55. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.] Chapter 7*, Unit.7.20 (2013).

56. Dayton, J. B. & Piccolo, S. R. Classifying cancer genome aberrations by their mutually exclusive effects on transcription. *BMC Medical Genomics* **10**, 66 (2017).

57. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

58. Pritchard, C. C. et al. Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. *The Journal of molecular diagnostics: JMD* **16**, 56–67 (2014).

59. Landrum, M. J. et al. ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research* **44**, D862–D868 (2016).

60. Landrum, M. J. et al. ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research* **44**, D862–D868 (2016).

61. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. deconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology* **17**, (2016).

62. Goldman, M. et al. The UCSC Cancer Genomics Browser: Update 2015. *Nucleic Acids Research* **43**, D812–D817 (2015).
63. Price, M. E. et al. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics & Chromatin* **6**, 4 (2013).

64. Loo, M. van der. Extremevalues: An R-package for distribution-based outlier detection. (2017).

65. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology* **12**, R41 (2011).

66. Nielsen, T. O. *et al.* A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* **16**, 5222–5232 (2010).

67. Netanely, D., Avraham, A., Ben-Baruch, A., Evron, E. & Shamir, R. Expression and methylation patterns partition luminal-A breast tumors into distinct prognostic subgroups. *Breast Cancer Research* **18**, 74 (2016).

68. Rahman, M. *et al.* Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. *Bioinformatics (Oxford, England)* **31**, 3666–3672 (2015).

69. Sioutos, N. *et al.* NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics* **40**, 30–43 (2007).

70. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2016).

71. Wickham, H., Hester, J. & Francois, R. **Readr**: Read Tabular Data. (2016).

72. Wickham, H. & Francois, R. **Dplyr**: A Grammar of Data Manipulation. (2016).

73. Wickham, H. **Ggplot2: Elegant Graphics for Data Analysis.** (Springer-Verlag New York, 2009).

74. Wickham, H. & Henry, L. **Tidyr**: Easily Tidy Data with ’spread()’ and ’gather()’ Functions. (2018).

75. Wickham, H. **Reshaping Data with the reshape Package.** *Journal of Statistical Software* **21**, 1–20 (2007).

76. Slowikowski, K. **Ggrepel**: Automatically Position Non-Overlapping Text Labels with ’ggplot2’. (2018).

77. Wilke, C. O. **Cowplot: Streamlined Plot Theme and Plot Annotations for ’ggplot2’.** (2017).

78. Dowle, M. & Srinivasan, A. **Data.Table**: Extension of ’data.Frame’. (2018).

79. Gehlenborg, N. **UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets.** (2017).
80. Team, T. B. D. BSgenome.Hsapiens.UCSC.Hg38: Full genome sequences for Homo sapiens (UCSC version hg38). (2015).

81. Huber, W. et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods 12, 115–121 (2015).

82. Krijthe, J. H. Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation. (2015).

83. van der Maaten, L. & Hinton, G. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9, 2579–2605 (2008).

84. van der Maaten, L. Accelerating t-SNE using Tree-Based Algorithms. Journal of Machine Learning Research 15, 3221–3245 (2014).

85. Hayden, D., Lazar, P., Schoenfeld, D. & Inflammation and the Host Response to Injury Investigators. Assessing statistical significance in microarray experiments using the distance between microarrays. PloS One 4, e5838 (2009).

86. Knudson, A. G. Mutation and Cancer: Statistical Study of Retinoblastoma. Proceedings of the National Academy of Sciences of the United States of America 68, 820–823 (1971).

87. Farmer, H. et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature 434, 917–921 (2005).

88. Bryant, H. E. et al. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. Nature 434, 913–917 (2005).

89. Tutt, A. et al. Carboplatin in BRCA1/2 -mutated and triple-negative breast cancer BRCAness subgroups: The TNT Trial. Nature Medicine 24, 628–637 (2018).

90. Zhao, W. et al. Promotion of RAD51-mediated homologous DNA pairing by BRCA1-BARD1. Nature 550, 360–365 (2017).

91. Couch, F. J. et al. Inherited mutations in 17 breast cancer susceptibility genes among a large triple-negative breast cancer cohort unselected for family history of breast cancer. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology 33, 304–311 (2015).

92. Shimelis, H. et al. Triple-Negative Breast Cancer Risk Genes Identified by Multigene Hereditary Cancer Panel Testing. Journal of the National Cancer Institute (2018). doi:10.1093/jnci/djy106
93. Crook, T. et al. P53 mutation with frequent novel codons but not a mutator phenotype in BRCA1- and BRCA2-associated breast tumours. *Oncogene* **17**, 1681–1689 (1998).

94. Greenblatt, M. S., Chappuis, P. O., Bond, J. P., Hamel, N. & Foulkes, W. D. TP53 mutations in breast cancer associated with BRCA1 or BRCA2 germ-line mutations: Distinctive spectrum and structural distribution. *Cancer Research* **61**, 4092–4097 (2001).