nature genetics

A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer

Paz Polak[1–3,21], Jaegil Kim[1,21], Lior Z Braunstein[1,4,21], Rosa Karlic[5], Nicholas J Haradhavala[1,2], Grace Tiao[1], Daniel Rosebrock[1], Dimitri Livitz[1], Kirsten Kübler[1–3], Kent W Mouw[3,6], Atanas Kamburov[1–3], Yosef E Maruvka[1–3], Ignaty Leshchiner[1], Eric S Lander[1,7,8], Todd R Golub[1,3,9], Aviad Zick[10], Alexandre Orthwein[11], Michael S Lawrence[1–3], Rajbir N Batra[12–14], Carlos Caldas[12–14], Daniel A Haber[2,3], Peter W Laird[15], Hui Shen[15], Leif W Ellisen[2,3], Alan D D'Andrea[16,17], Stephen J Chanock[18], William D Foulkes[11,19,21] & Gad Getz[1–3,20,21]

Biallelic inactivation of *BRCA1* or *BRCA2* is associated with a pattern of genome-wide mutations known as signature 3. By analyzing ~1,000 breast cancer samples, we confirmed this association and established that germline nonsense and frameshift variants in *PALB2*, but not in *ATM* or *CHEK2*, can also give rise to the same signature. We were able to accurately classify missense *BRCA1* or *BRCA2* variants known to impair homologous recombination (HR) on the basis of this signature. Finally, we show that epigenetic silencing of *RAD51C* and *BRCA1* by promoter methylation is strongly associated with signature 3 and, in our data set, was highly enriched in basal-like breast cancers in young individuals of African descent.

Breast cancer is the most common noncutaneous malignancy in women, and approximately 10–15% of cases are associated with hereditary syndromes of defective DNA repair[1]. The most common forms of hereditary breast and ovarian cancer are associated with pathogenic germline variants in two HR repair genes, *BRCA1* and *BRCA2* (ref. 2). HR is an evolutionarily conserved pathway that coordinates high-fidelity repair of double-stranded DNA breaks (DSBs), and functions primarily during the late S and G2 phases of the cell cycle to exploit the intact sister chromatid as a template for error-free repair[3]. Although germline alterations that abrogate the tumor-suppressor functions of *BRCA1* and *BRCA2* have a role in breast cancer pathogenesis, comprehensive sequencing studies have demonstrated that somatic mutations and epigenetic events in these genes can also drive tumorigenesis[4–7]. Deficiency of HR (HRD) leads to cellular dependence on alternative, error-prone DNA-repair pathways, thus yielding characteristic genomic alterations, higher overall mutational rates, and unique dependencies that can be used for therapeutic targeting[8].

Analyses of tumor-derived genome sequences have shown that the loss of *BRCA1* or *BRCA2* yields a distinct pattern of base-substitution mutations termed signature 3 (refs. 9,10). This pattern, among others, can be discerned via non-negative matrix factorization (NMF), a technique used to identify recurring patterns (i.e., signatures) in the spectra of mutations from a set of tumors and to estimate the contributions of these signatures to the mutational landscape of the individual tumors. The etiology of some signatures has been revealed through association of their activities with additional data[10–13]. Although recent efforts have shown a significant association between the loss of *BRCA1/2* and signature 3, several tumors show high levels of signature 3 without discernible alterations in HR-pathway genes[9,10]. Furthermore, signature 3 has not been extensively correlated with other genetic and epigenetic events in HR genes. A comprehensive understanding of the link between the HR pathway and signature 3 is of clinical relevance given the implications for underlying cancer risk and treatment selection.

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Massachusetts General Hospital Center for Cancer Research, Charlestown, Massachusetts, USA. [3]Harvard Medical School, Boston, Massachusetts, USA. [4]Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [5]Bioinformatics Group, Division of Molecular Biology, Department of Biology, Faculty of Science, University of Zagreb, Zagreb, Croatia. [6]Department of Radiation Oncology, Brigham & Women's Hospital and Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [7]Department of Biology, MIT, Cambridge, Massachusetts, USA. [8]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [9]Departments of Medical Oncology, Pathology, and Pediatric Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [10]Department of Oncology, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [11]Department of Oncology, McGill University, Montreal, Quebec, Canada. [12]Cancer Research UK Cambridge Institute, University of Cambridge Li Ka Shing Centre, Cambridge, UK. [13]Cancer Research UK Cambridge Centre, Cambridge, UK. [14]Department of Oncology, University of Cambridge Hutchison–MRC Research Centre, Cambridge Biomedical Campus, Cambridge, UK. [15]Van Andel Research Institute, Grand Rapids, Michigan, USA. [16]Center for DNA Damage and Repair, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [17]Ludwig Center at Harvard, Harvard Medical School, Boston, Massachusetts, USA. [18]Division of Cancer Epidemiology and Genetics, US National Cancer Institute, Bethesda, Maryland, USA. [19]Department of Human Genetics, Lady Davis Institute for Medical Research and Research Institute McGill University Health Centre, McGill University, Montreal, Quebec, Canada. [20]Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, USA. [21]These authors contributed equally to this work. Correspondence should be addressed to G.G. (gadgetz@broadinstitute.org).

Received 16 December 2016; accepted 21 July 2017; published online 21 August 2017; doi:10.1038/ng.3934

Current clinical practice in testing for breast cancer susceptibility relies largely on the identification of known pathogenic variants in the germline sequence of *BRCA1/2* and a limited number of related genes[14,15]. Such germline testing may uncover pathogenic variants that are likely to contribute to HRD in the associated breast tumor. As the presence of HRD may expose specific therapeutic vulnerabilities, we wanted to assess whether the integration of germline and somatic data can improve the detection of HRD. Somatic events, such as loss of heterozygosity, epigenetic silencing, and other mechanisms that can decrease gene expression, can serve as the 'second hit' in these genes[16]. Moreover, integration of somatic data can help researchers interpret poorly defined germline variants[17]. Here we hypothesize that signature 3 could represent a more reliable readout of HR status than germline sequencing alone and could lead to improved prediction of breast cancer risk. Moreover, signature 3 may aid in the classification of variants that have yet to be functionally characterized. We therefore sought to identify the mechanisms by which signature 3 could arise in the absence of pathogenic germline alterations in *BRCA1/2*. Specifically, we analyzed the association between signature 3 and multidimensional events (germline variants, somatic mutations, epigenetic silencing, and allelic imbalance) in HR-pathway components aside from *BRCA1/2*.

RESULTS

A distinct mutational signature is associated with biallelic inactivation of *BRCA1/2* and epigenetic silencing of *BRCA1*

We first characterized the underlying landscape of genomic lesions that operate in breast cancer by carrying out mutational signature analysis of 995 invasive breast carcinoma samples from The Cancer Genome Atlas (TCGA)[4]. We stratified single-nucleotide variants (SNVs) by base substitutions in 96 possible trinucleotide contexts (**Fig. 1**) and applied our method SignatureAnalyzer[12,13], which uses a Bayesian derivative of NMF to infer the number of operating signatures as well as the activity of each signature in each tumor (Online Methods).

We identified seven distinct mutational signatures across the 995 TCGA breast cancers, largely corresponding to previous analyses of this series[10] (**Supplementary Fig. 1** and Online Methods). Of these, three signatures arose in a single case each, and we excluded them from subsequent analyses (Online Methods). This yielded four recurrent signatures across 992 samples: (i) signature 1 (C>T at CpG sites), (ii) an APOBEC-related signature (a combination of signature 2 and signature 13), (iii) a signature associated with microsatellite instability (signature 6), and (iv) a largely uniform signature, similar to the previously reported signature 3, that was associated with mutations in *BRCA1/2* (refs. 9,10,18) (**Fig. 1**, **Supplementary Fig. 2**).

Over 10% (100/992) of the samples harbored clinically relevant germline or somatic frameshift/nonsense variants in *BRCA1/2*, well-characterized pathogenic missense *BRCA1/2* germline variants[19], and/or epigenetic silencing of the *BRCA1* promoter[16]. As previously demonstrated, a variety of *BRCA1/2* inactivating variants were associated with signature 3 (refs. 9,10,18) (**Fig. 2**, **Supplementary Fig. 3**). Fifty samples with somatic or germline nonsense/frameshift variants in *BRCA1*/2 showed an approximately sixfold increase in levels of signature 3 compared with those in samples without *BRCA1/2* alterations ($P < 0.001$, rank sum test; **Fig. 3a**), consistent with the enrichment of

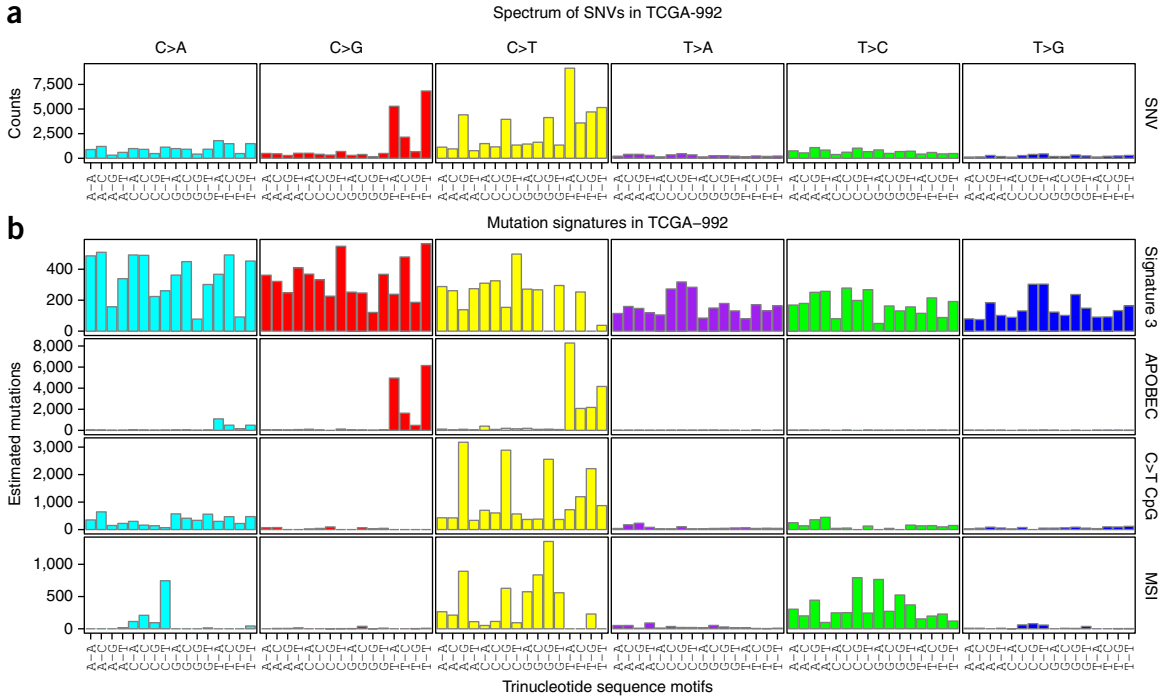

Figure 1 Characterization of four distinct mutational signatures in breast cancer. (**a**,**b**) Each color represents one of the six potential base substitutions, with each substitution further stratified by the adjacent 5′ and 3′ flanking nucleotides. (**a**) We used a Bayesian NMF approach (Online Methods) to decompose the overall mutational spectrum across a cohort of 992 breast cancer samples into distinct mutational signatures (SNVs). (**b**) We compared the pattern of each signature with a database of known signatures and their etiologies (signature 3, C>T_CpG, APOBEC, and microsatellite instability (MSI)). A signature characterized by C>G transversions and C>T transitions at TC[A/T] motifs (in which the altered cytosine is flanked by a 5′ thymine and a 3′ adenine or thymine) is probably attributable to mutagenesis via endogenous APOBEC machinery (APOBEC; COSMIC signatures 2 and 13). A second signature consisting primarily of C>T transitions at CpG dinucleotides has been described among all tumor types and is thought to arise from spontaneous 5-methyl-cytosine deamination (C>T CpG; COSMIC signature 1). The third signature resembles previously characterized signatures identified in samples known to have MSI (COSMIC signature 6). A recurrent signature in the series closely approximates signature 3 and is characterized by mutations in all 96 nucleotide contexts, with an increased rate among C:G base pairs as compared with that among A:T.

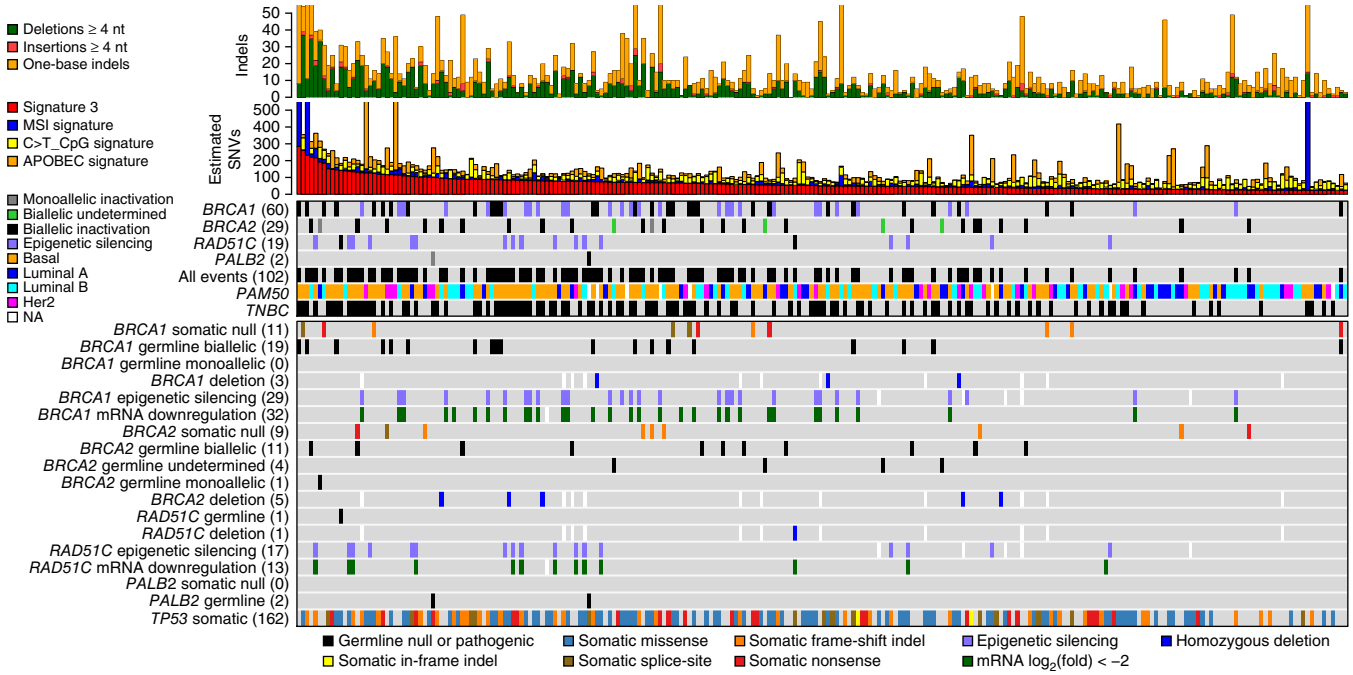

Figure 2 Overall mutation rates, mutational signature contributions, and clinicopathological features per patient as sorted by descending signature 3 activity. The figure shows the 250 samples with the highest signature 3 activity; **Supplementary Figure 3** represents the remaining 742 samples. Each column represents a tumor sample. The overall insertion/deletion (indel) burden is shown at the top. The "Estimated SNVs" graph shows the absolute mutation burden per sample, with the estimated contribution of each of the four mutational signatures designated by color-coding according to the key (samples are arranged in descending order of signature 3 activity). The third plot from the top shows sample-level annotations with regard to lesions in known HR-pathway genes, along with *PAM50* intrinsic subtype data, and triple-negative status determined as in ref. 4. The bottom plot shows genetic lesions by type in select HR genes, with mutation type indicated by color-coding according to the key.

signature 3 reported in a recent data set[9] (**Fig. 3b**). Aside from SNVs, epigenetic silencing of *BRCA1* (n = 32) and homozygous deletions of *BRCA1/2* (n = 8) were also associated with signature 3 (**Fig. 3a**). Moreover, 18 samples that harbored the lower-risk truncating *BRCA2* variant p.Lys3326Ter[20] (c.9976A>T, chr13:g.32972626A>T (hg19);NM_000059.3) did not show elevated levels of signature 3 compared with those in samples without inactivating *BRCA1/2* alterations (median of 13.4 signature 3 mutations per sample; P = 0.4; data not shown).

A single functional copy of *BRCA1/2* is generally sufficient to maintain normal HR function, whereas biallelic inactivation of *BRCA1/2* contributes to tumorigenesis via several mechanisms, including genomic instability that arises from HRD[21]. To investigate whether monoallelic loss of *BRCA1/2* is associated with signature 3 and whether HRD (biallelic loss) is necessary for signature 3, we analyzed loss of heterozygosity (LOH) at the *BRCA1/2* loci (Online Methods), as these are the most common second-hit events[22,23]. We identified 31 samples with pathogenic germline variants in *BRCA1/2* associated with LOH of the wild-type allele (**Supplementary Fig. 4**). All samples but one showed signature 3 activity in the top quartile for the series (**Fig. 2**, **Supplementary Fig. 4**). Similarly, all 18 samples with LOH that also harbored a somatic frameshift/nonsense/splice-site *BRCA1/2* variant had high signature 3 activity (**Supplementary Fig. 4**).

In contrast, monoallelic germline inactivation of *BRCA1/2* was not associated with signature 3 (**Supplementary Fig. 4**), in agreement with recent reports[9]. Of five samples that harbored pathogenic germline variants without loss of the intact allele, four showed no increase in signature 3 activity (P = 0.42). Two samples with a somatic truncating mutation and without evidence of biallelic inactivation also showed no signature 3 elevation (**Supplementary Fig. 4**). RNA-seq data for these five cases had very low coverage of *BRCA1/2* at the variant site (≤2 reads) and could not be used to examine expression of the mutant alleles compared with that of the wild type alleles. In addition, samples with LOH of *BRCA1/2* that retained an allele with a benign silent germline variant did not show strong enrichment of signature 3 (**Supplementary Fig. 4**; median of 12.9 signature-3-associated mutations, in contrast to 71 where LOH retained a truncating event). These findings suggest that signature 3 may be a reliable readout of biallelic, and not monoallelic, loss of *BRCA1/2* and/or of HRD.

Despite the association between *BRCA1/2* inactivation (germline, somatic, or epigenetic silencing) and signature 3 (88 of 100 *BRCA1/2* events were in the top quartile of signature 3 activity), most samples in the top quartile (159 of 247) did not have such events, which suggests that other lesions might contribute to signature 3 activity. Clinical assays for the assessment of breast cancer risk include gene panels of varying breadth, nearly all of which include a variety of DNA-repair genes. As discussed above, numerous germline variants of *BRCA1/2* confer high cancer risk; however, deleterious variants in several other genes confer 'moderate' risk[15] (**Supplementary Fig. 5**). Therefore, we sought to determine whether alterations in the latter group of genes are associated with signature 3 (**Fig. 3c**). In particular, we focused on germline events in which the intact allele was lost (**Supplementary Table 1**, **Supplementary Fig. 6**). In the following sections, we consider the association between deleterious variants in these genes and signature 3.

Signature 3 is associated with lesions in established breast cancer gene *PALB2*

In search of additional germline variants that could be associated with signature 3, we examined all 992 cases for any deleterious

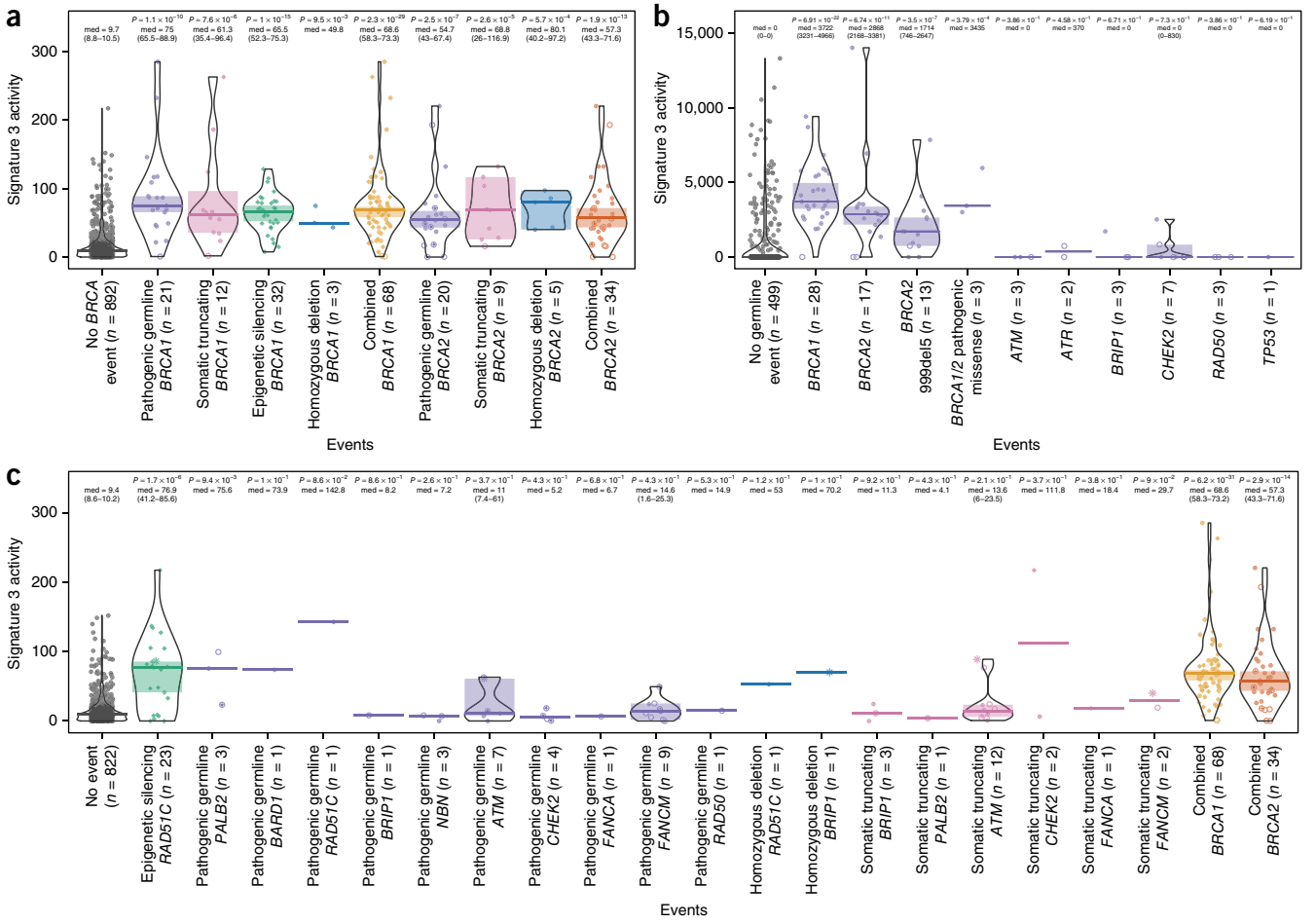

Figure 3 Signature 3 activity in tumors with somatic, germline, and epigenetic alterations in HR-pathway genes. (**a**) Signature 3 activity is elevated in tumors with somatic, germline, and epigenetic alterations in *BRCA1/2*. The plot shows the number of mutations attributable to signature 3 in tumors stratified by germline, somatic, or epigenetic events in *BRCA1/2*. Pathogenic germline events included those described by ClinVar, in addition to frameshift and nonsense variants. We compared the median signature count (med) among each group with that of samples that lacked any discernible *BRCA1/2* alterations ("no *BRCA* event"). Two samples harbored a somatic and a germline truncating event in *BRCA1/2* and are included in both the somatic and germline categories. (**b**) Signature 3 activity in subjects with germline alterations as previously denoted[9]. *P* value comparisons are to samples with no germline event. (**c**) Signature 3 activity is associated with somatic, germline, and epigenetic alterations in HR components beyond *BRCA1/2*. The plot shows signature 3 activity for samples that lacked any of the noted alterations ("no event") and those with any genetic or epigenetic event in the HR pathway. Samples with an event in *BRCA1/2* are noted by an asterisk. In all panels, solid circles represent samples with evidence for biallelic inactivation, open circles represent samples with monoallelic lesions (i.e., without evidence of biallelic inactivation), open circles with crosses represent samples with pathogenic germline mutation and LOH of the gene but no clear determination of which of the two alleles was lost, and diamonds indicate samples with epigenetic silencing. Each symbol represents an individual sample. Horizontal lines represent the median. Colors represent the type of event: purple, pathogenic germline; pink, somatic truncating; green, epigenetic silencing; blue, homozygous deletion; yellow, all combined BRCA1 events; orange, all combined BRCA2 events; black, no event. Black outlines represent the kernel probability density of the data in groups where the number of samples is ≥5. *P* values were determined by Wilcoxon rank-sum test.

alterations in other known genes of the HR pathway, including *PALB2*, a binding partner and nuclear localizer of BRCA2 (**Supplementary Fig. 6**). In recent studies, germline truncating variants in *PALB2* have been associated with a relative risk for breast cancer of between 3.4 and 6.6 (refs. 15,24,25). Three samples in our cohort harbored germline nonsense/frameshift variants in *PALB2*, and all exhibited elevated signature 3 activity (**Fig. 3c**). We could not validate this association in a recent report of 560 breast cancers[9] because of a lack of these germline events. In agreement with our results, however, in a recent pancreatic study[26,27], two cases with *PALB2* germline truncating variants had high levels of signature 3.

RAD51C methylation as a key alteration underlying deficient HR repair in basal-like breast carcinoma

RAD51 is involved in DNA repair by HR, promoting DNA-strand invasion and homology search[14]. This process requires RAD51 in complex with BRCA2 and, to a lesser extent, PALB2 (ref. 28), both of which can promote loading of RAD51 onto single-stranded DNA. There are several *RAD51*-related genes, including *RAD51B*, *RAD51C*, *RAD51D*, *DMC1*, *XRCC2*, and *XRCC3*. RAD51C facilitates the accumulation of RAD51 at sites of DNA damage by complexing with several RAD51-like proteins (**Fig. 4a**).

It is generally thought that germline truncating variants in *RAD51C* do not confer an increased risk of breast cancer[29], although

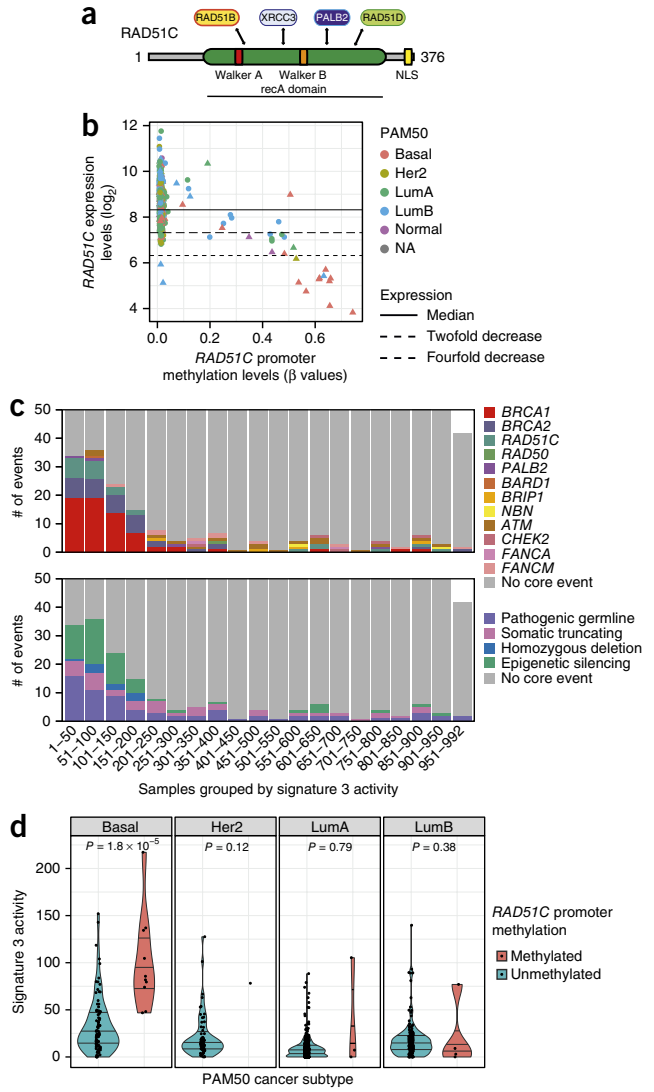

Figure 4 The association of *RAD51C*-promoter methylation with elevated signature 3 activity in basal-like tumors. (**a**) RAD51C architecture and known interacting proteins that promote HR. (**b**) *RAD51C* expression versus promoter methylation levels (methylation values higher than 0.2 were considered for epigenetic silencing). Each symbol represents an individual sample. Triangles denote samples with high signature 3 activity (top quartile). (**c**) Elevated signature 3 activity is associated with alterations of multiple components of the HR pathway, and a diversity of alteration types. Both plots show the number of signature-3-associated mutations in the samples studied, in descending order in bins of 50. The top plot shows alterations among various HR-pathway genes, and the bottom plot shows the types of alterations in these genes, demonstrating the enrichment of events among the samples with the highest signature 3 activity. (**d**) Signature 3 activity was significantly associated with *RAD51C*-promoter methylation among basal-like breast cancers, but not among other intrinsic subtypes.

widespread panel-based genetic testing has revealed deleterious *RAD51C* variants among 0.3% of people with triple-negative breast cancer[30] and in high-risk breast pedigrees[31]. In our data, one case showed biallelic inactivation of *RAD51C* (a germline variant with LOH that deleted the intact allele), and another case presented somatic homozygous deletion. Both tumors had high levels of signature 3 (**Figs. 2** and **3c**), which suggests that inactivation of *RAD51C* can lead to HRD.

As *BRCA1*-promoter hypermethylation can promote breast cancer development, we investigated whether such events also occur in other HR genes. Indeed, *RAD51C*-promoter methylation correlated with downregulated gene expression in our series (**Fig. 4b**). Of 23 cases of *RAD51C*-promoter methylation, 19 showed at least a twofold decrease in mRNA expression compared with the median across all samples (**Fig. 4b**). Seventeen of the 23 samples were in the top quartile of signature 3 activity, and all but one had no *BRCA1/2* events (the one with a *BRCA1/2* event had elevated *RAD51C* expression). Thus, 16 cases of *RAD51C* epigenetic silencing explained 10% (16/159) of the top quartile without *BRCA1/2* events (**Fig. 4c**). Moreover, the number of samples with high signature 3 activity that harbored epigenetic events in *BRCA1* and *RAD51C* was comparable to the number of germline events in *BRCA1/2*, which suggests a similar contribution of germline and epigenetic events to carcinogenesis (**Fig. 4c**).

Of note, 52% (12/23) of samples with *RAD51C* silencing had the basal-like expression pattern, and all of these basal-like cases were among the 17 *RAD51C*-silenced samples in the top signature 3 quartile (70%; $P = 0.003$, Fisher's exact test). Examination of an independent series[32,33] with methylation and expression data showed a similar correlation between promoter methylation and expression of *RAD51C* (**Supplementary Fig. 7**). In agreement with these findings, cases with *RAD51C*-promoter methylation and low *RAD51C* expression levels showed high enrichment with basal-like tumors (**Fig. 4b**, **Supplementary Fig. 7**). Basal-like expression is characteristic of breast cancers with *BRCA1* defects[34] and is seen in 81% (17/21) of cases with germline *BRCA1* deleterious variants and 81% (25/32) of cases with *BRCA1*-promoter hypermethylation[35,36]; we also observed this last association in the independent series (**Supplementary Fig. 7**). In contrast, only 10% (2/20) of tumors with *BRCA2* germline deleterious variants were basal-like, which suggests that inactivation of *RAD51C* may be biologically closer to *BRCA1* defects than to *BRCA2* inactivation. *RAD51C*-promoter methylation was associated with increased levels of signature 3 only in basal-like tumors (**Fig. 4d**), consistent with the strongly reduced expression in this subtype, which suggests that *RAD51C*-promoter methylation is involved in HRD mainly in the context of basal-like tumors.

No association of signature 3 with mutations in DNA-damage signaling pathway genes

The role of *BRCA1/2* in PARP-inhibitor susceptibility is established. However, the therapeutic significance of deleterious variants in HR components upstream of BRCA1/2 (**Supplementary Fig. 5**) that are also implicated in DNA-damage signaling and DSB detection and that increase the risk of breast cancer[15] (relative risk, 2–3), such as ATM, NBN, and CHEK2, is less clear. In our data set, we identified 17 germline pathogenic variants in six established or candidate breast-cancer-susceptibility genes (7 in *ATM*; 4 in *CHEK2*; 3 in *NBN*; and 1 in *BARD1*, *BRIP1*, and *RAD50*; a discussion is provided in **Supplementary Note 1**). Overall, germline pathogenic variants in *ATM* and *CHEK2* were not associated with a high level of signature 3 (**Supplementary Table 1**, **Fig. 3c**), and this result was replicated in a previously published data set[9] (**Fig. 3b**). The single case with a *BARD1* truncating variant and LOH (loss of the wild-type allele) had a high level of signature 3 (**Fig. 3c** and **Supplementary Fig. 6**).

Promoter-methylation events in *BRCA1* and *RAD51C* are enriched among young black individuals

We next sought to assess whether self-reported racial differences underlie various signature 3 etiologies. The largest racial groups in our series consisted of 699 white subjects and 142 black/African American

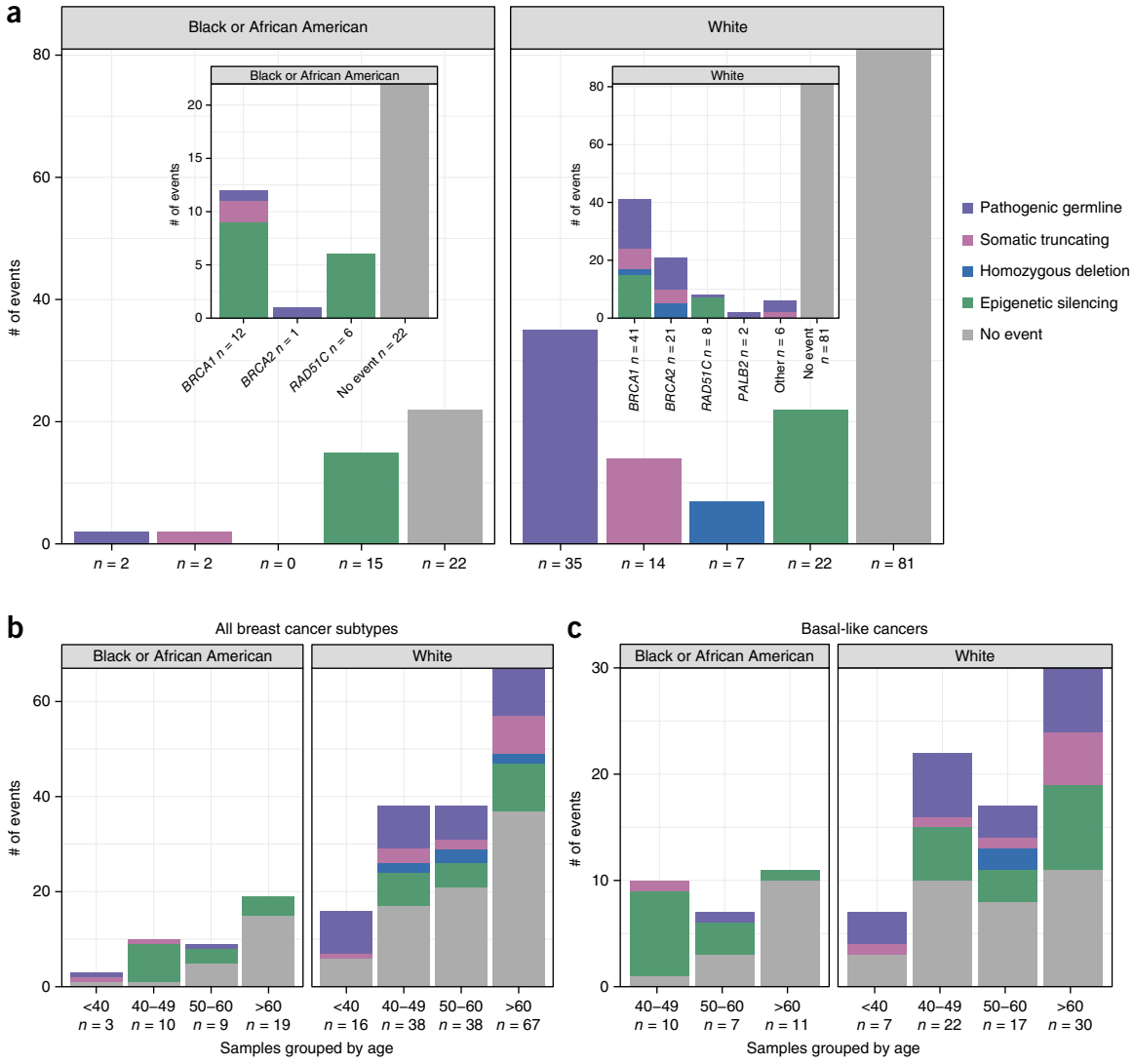

Figure 5 Analysis of genetic and epigenetic events in the top quartile of signature 3 activity ($n = 248$), focused on the largest racial subgroups in our cohort (white and African American). (**a**) The distribution of various types of alterations among subjects stratified by racial group, showing enrichment of promoter methylation in African American subjects. The insets show the distribution of various types of alterations in individual genes. (**b**,**c**) The enrichment of promoter methylation among African American subjects was most pronounced among women aged 40–49 when looking at both all breast cancer subtypes (**b**) and only basal-like cancers (**c**).

subjects (the remaining 181 included 57 Asian patients). We evaluated differences in the frequencies of genetic (germline and somatic) and epigenetic events that contribute to signature 3 between these racial groups, focusing on the cases in the top quartile of signature 3 activity (159 white and 41 black/African American subjects).

Among the 41 black subjects in the top quartile, 37% (15/41) showed promoter methylation in either *BRCA1* or *RAD51C*, 9.8% (4/41) harbored germline or somatic deleterious mutations in *BRCA1/2*, and the remaining 22 had no discernible lesion (**Fig. 5a**). In contrast, among the 159 white subjects in the top quartile, 13% (22/159) showed an epigenetic event, 32% (56/159) had germline events or somatic deleterious events, and the remaining 81 had no event (**Fig. 5a**; Freeman–Halton 3×2 exact test; $P = 3 \times 10^{-4}$ for white compared with black subjects). Reductions in levels of *RAD51C* and *BRCA1* mRNA in cases with promoter methylation were similar among white and black individuals, suggesting similar function effects.

Among premenopausal women, African Americans are at higher risk of developing basal-like breast cancers than their non-Hispanic white counterparts[37]. Given the association between HRD and basal-like biology, as well as the observed differences between racial groups in the prevalence of signature-3-associated genetic and epigenetic events, we investigated differences between age subgroups (**Fig. 5b**) and within the basal-like subtype (**Fig. 5c**). The most pronounced differences arose among women aged 40–49 years. Of black women in this age group, 80% (8/10) showed promoter methylation (6 *BRCA1* and 2 *RAD51C*), in contrast to only 18% (7/38) of white women. Conversely, genetic events were enriched among white women (36%; 14/38), whereas only one black woman in this age group had a genetic alteration (1/10). This suggests that the mechanism of HRD differs between white people, in whom genetic mutations predominate, and black people, in whom promoter methylation is more frequent ($P = 0.0009$). We further examined

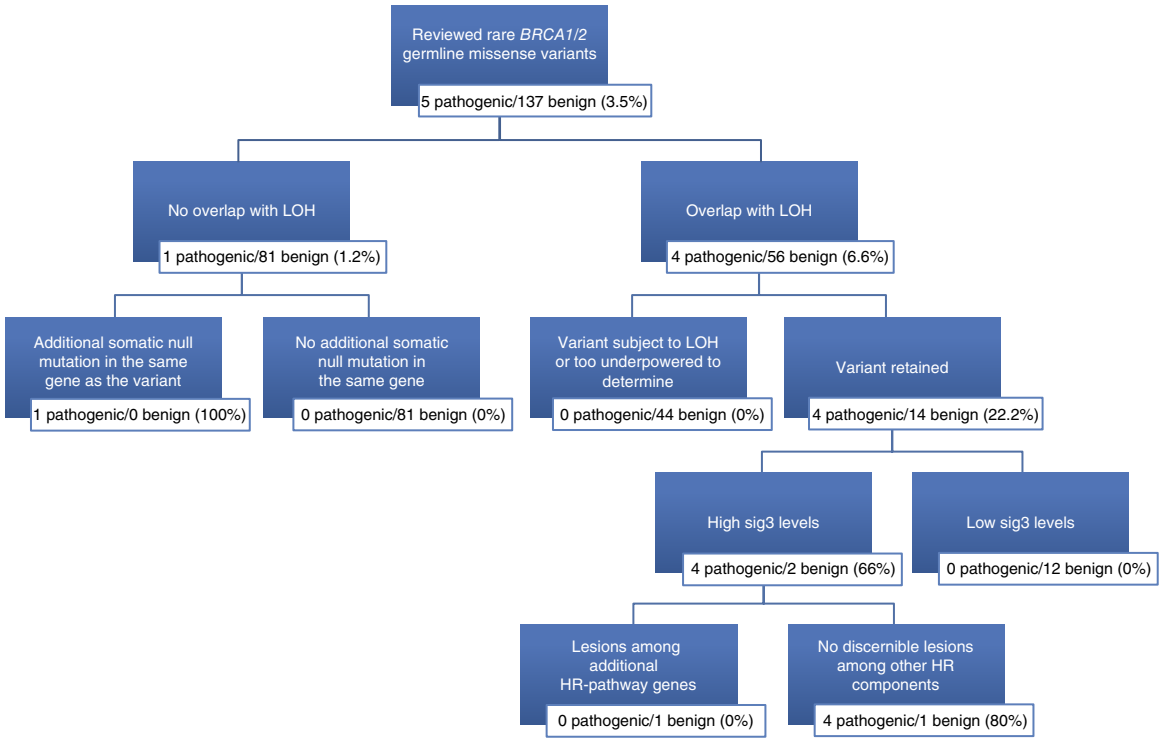

Figure 6 A framework for enhancing the classification of *BRCA1/2* germline missense variants using signature 3 (sig3) and biallelic inactivation. Beginning with all rare *BRCA1/2* missense germline variants ($n = 137$; 3.5% of randomly selected variants in the TCGA breast cancer data set are pathogenic), we dichotomized the group on the basis of the presence of LOH at the *BRCA1/2* locus. Those without LOH ($n = 82$) included only one sample with a pathogenic allele (i.e., 1.2% probability of pathogenicity), although this sample also harbored an additional variant with the potential for biallelic inactivation. Conversely, among those with LOH ($n = 60$; 6.6% of variants were pathogenic), samples were further sorted on the basis of allelic imbalance (22.2% of variants were pathogenic), elevated signature 3 activity (66.6% probability of pathogenicity), and the absence of lesions among other HR-pathway components

whether this difference also arises in basal-like individuals aged 40–49 and found a similar disparity (**Fig. 5c**; $P = 0.0065$).

Signature 3 activity correlates with the pathogenic potential of rare germline variants

We next sought to evaluate whether signature 3 can be used to characterize rare missense SNVs in *BRCA1/2*. Understanding the pathogenicity of rare germline variants of *BRCA1/2* has become the focus of several international consortia[38,39] and has led to the assignment of consensus annotations to rare variants into five classes: benign, likely benign, variants of uncertain significance (VUSs), likely pathogenic, and pathogenic[40]. Among the entire series of 992 cases, 262 (26.3%) harbored at least one rare (prevalence < 0.1%) missense SNV in *BRCA1/2*.

As discussed above, the biallelic inactivation of *BRCA1/2*, particularly via an inherited pathogenic variant and LOH, is considered a hallmark of HRD. As shown above, nearly all samples with pathogenic *BRCA1/2* germline variants and loss of the intact allele had high levels of signature 3, whereas those that harbored pathogenic variants without LOH had no elevation of signature 3. Conversely, samples with LOH that otherwise retained an allele harboring a silent germline variant were not associated with a significant increase in signature 3 (**Supplementary Fig. 4**), which suggests that the presence of LOH alone is not enough to accurately identify samples with HRD. Signature 3 in conjunction with LOH, however, could represent a robust readout of HRD and of the pathogenic potential of indeterminate variants. Therefore, we sought to apply a 'biallelic inactivation' model with high signature 3 levels to potentially discriminate between pathogenic and benign variants. We estimated the likely pathogenicity of a variant by using two methodologies based on the biallelic inactivation model for HRD (Online Methods).

In the 262 cases discussed above, we observed 143 distinct missense germline SNVs in *BRCA1/2*. Seventy-four were present in more than one sample, of which eight also exhibited LOH (loss of the intact allele) in at least two tumors (**Supplementary Table 2**). Of these eight, one is considered pathogenic, and seven are thought to be benign[19]. The pathogenic variant, BRCA1 p.Cys61Gly (HGSV nomenclature is shown in **Supplementary Table 2**), was present in four samples, all of which exhibited high signature 3 activity along with biallelic inactivation (three with LOH, and one with an additional somatic truncating variant; $P = 0.004$; **Supplementary Table 2**). p.Cys61Gly in BRCA1 is a founder variant among Poles and disrupts the RING-domain interaction between BRCA1 and BARD1 (refs. 41,42). Here we identified it via a largely unbiased approach by seeking rare germline missense SNVs with concomitant LOH, without familial, pathological, or other clinically relevant data. The remaining seven germline variants, previously described as benign, were not associated with an increase in signature 3 activity.

We hypothesized that LOH status in tandem with signature 3 might further inform on the clinical implications of poorly characterized *BRCA1/2* germline variants. To test this approach, we used the set of germline missense SNVs in our data for which a functional implication had been classified by the ENIGMA consortium ($n = 142$): 5 as pathogenic variants (class 5 in ClinVar), and the remaining 137

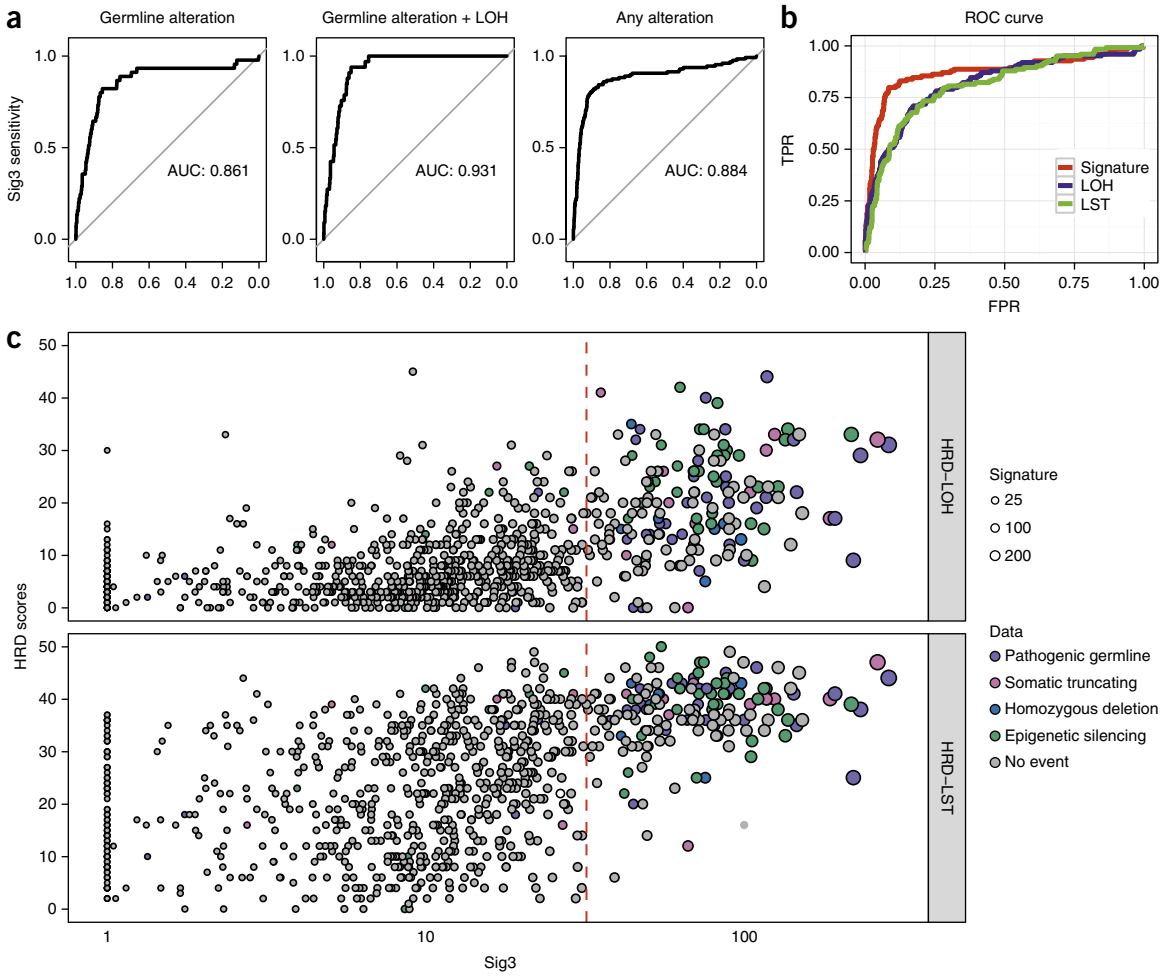

Figure 7 Prediction accuracy of signature 3 (sig3) relative to established rearrangement-based HRD scores. (**a**) Receiver operating characteristic (ROC) curves for alterations among *BRCA1*, *BRCA2*, *RAD51C*, and *PALB2* predicted on the basis of signature 3 levels. The panels show analyses based on (i) all pathogenic germline mutations, (ii) pathogenic germline mutations with LOH in the tumor, and (iii) pathogenic germline, loss-of-function somatic, or epigenetic alterations. (**b**) ROC curves comparing signature 3 performance to LOH and LST scores for identifying samples with the HR-associated lesions mentioned in **a**. (**c**) Signature 3 levels plotted against HRD–LOH and HRD–LST scores. Each symbol represents an individual sample.

as benign (class 1)[39] (**Fig. 6**). Therefore, in our data set the fraction of pathogenic missense variants was 3.5% (95% confidence interval (CI), (1%, 8%)). However, among samples with germline variants in which allelic imbalance favored the alternate allele, 22% harbored pathogenic variants (4/18; CI, (6%, 47%); $P < 0.001$ by hypergeometric test). Including the even more stringent requirement that a variant exhibit both allelic imbalance and high signature 3 activity (top quartile) yielded a set in which 66% of samples harbored pathogenic alleles (4/6; CI, (22%, 99%); $P < 4.1 \times 10^{-6}$; **Fig. 6**); one of the two apparently benign variants also harbored *RAD51C* hypermethylation, which can underlie high signature 3 activity. Thus the combination of LOH and signature 3 activity considerably improves the likelihood of pathogenicity for a given allele. Conversely, among 16 samples with an excess of the alternative allele but low levels of signature 3, all *BRCA1/2* variants were benign, which suggests that this method is not only sensitive but also highly specific. Moreover, in an independent data set of whole genomes[9], we observed that two pathogenic variants (class 5) and one likely pathogenic (class 4) variant (**Fig. 3b**) had high levels of signature 3 concomitant with an LOH event that favored retention of the mutant allele[9], thereby demonstrating the sensitivity of this approach, independent of method or platform.

By considering that either LOH or a second deleterious mutation results in biallelic inactivation, we were able to use our approach to capture the known pathogenic *BRCA1/2* mutations among the series and accurately classify *BRCA1/2* VUSs. Overall, 12 unclassified missense variants fulfilled the criteria (**Supplementary Note 1**). Downstream, only three potentially pathogenic rare missense mutations had supportive evidence (**Supplementary Table 3**). Thus, in our data set of ~1,000 breast cancers, only eight rare missense SNVs (of which five were previously known) were likely to have contributed to the development of HR-deficient tumors. We applied a similar approach to somatic events, and the results obtained suggested that four somatic missense mutations (all in *BRCA1*) may disrupt critical functions in the HR pathway (**Supplementary Note 1**).

Signature 3 is a robust and independent biomarker of HRD

Finally, because signature 3 could be associated with the degree of HRD, we sought to establish a clinically applicable threshold for binary phenotypic classification. Such a task is difficult because there is no standard with which to unequivocally establish whether HRD is present. However, germline events in *BRCA1/2* and *PALB2* are known contributors to HRD. To standardize a measure to estimate the predictive

power of signature 3 for the identification of samples with events in HR-pathway genes, we calculated the area under the receiver operating characteristic curve (AUC), using all cases with deleterious alterations in *BRCA1/2*, *PALB2*, and *RAD51C* as a ground truth positive set. The AUC for detecting deleterious alterations was 0.86, and this value increased to 0.93 when we considered only germline cases with a second hit as the ground truth set (**Fig. 7a**). A similar AUC for the classification of HRD events was recapitulated when we used whole genomes or only 8 Mb of sequence and applied a normalized version of signature 3. In addition, signature 3 levels were not correlated with purity in our TCGA series (**Supplementary Note 1**, **Supplementary Figs. 8–11**). Thus signature 3 activity may serve as a robust classifier of HR events across different platforms.

Assays for HRD are used to identify samples with underlying events beyond germline lesions in *BRCA1/2*, such as somatic events in HR genes or epigenetic silencing. The calculated AUC with all events was 0.89 (**Fig. 7a**)—slightly higher than that calculated on the basis of *BRCA1/2* germline events, but lower than the estimated AUC for germline events with LOH. These results were superior to those obtained with alternative approaches based on LOH[43] and large-scale transition[44] scores (referred to as HRD–LOH and HRD–LST, respectively) (**Fig. 7b**), largely because signature 3 is more sensitive for the identification of *BRCA2* and *RAD51C* events (**Supplementary Fig. 12**). We found that a threshold of 31 signature-3-associated mutations (chosen to maximize Youden's index[45]) allowed us to detect 82% of the events in our data set with HRD, and might be sufficient to identify another 11% of samples with a yet unknown defect in HR (**Supplementary Fig. 13**). Among samples with more than 31 mutations, we found a putatively etiologic lesion in roughly 50%. Most, however, also harbored high HRD–LOH and HRD–LST scores (**Fig. 7c**), suggesting elevated degrees of genomic instability. We observed that the optimal threshold for the classification of biallelic inactivation of *BRCA1/2* events was 37 signature-3-associated mutations (**Supplementary Fig. 14**, **Supplementary Note 1**). As clinical applications often assay a subset of genes, we conducted tests to determine how many genes would yield an accurate classification of subjects with high levels of signature 3 (as defined by exome sequencing) and found that the use of 1,600 genes led to an AUC of ~0.82 (**Supplementary Fig. 15**, **Supplementary Note 1**).

DISCUSSION

Our analysis represents an extended characterization of signature 3 (refs. 9,10,18) and its relationship to the functional underpinnings of the HR repair machinery. Using a multidimensional approach, we reliably detected this signature among samples with deleterious germline and somatic mutations in *BRCA1* and *BRCA2*. In agreement with published results[9], we demonstrated and quantified the association of biallelic inactivation of *BRCA1/2* with signature 3, and observed a lack of association in cases with monoallelic inactivation.

We further used signature 3 to identify a large subset of tumors (16%) with no discernible canonical HR defect, yet with a mutational landscape similar to that of samples with HRD. When we focused on this subset, we found that a small number of them harbored rare truncating variants in other components of the HR machinery.

Most important, we demonstrated that epigenetic silencing and somatic mutations in *RAD51C* have similar potential to abrogate HR function and yield the same characteristic mutational signature. The frequency of *RAD51C*-promoter methylation events was similar to that of *BRCA2* germline events and higher than that of somatic truncating *BRCA1* events in these tumors. *RAD51C* silencing events were even more frequent in basal-like breast cancers. Although *PALB2* is currently considered to be an important determinant gene for hereditary breast cancer[15], we found only 2 samples that harbored germline *PALB2* lesions among the top quartile of signature 3 activity, in contrast to 18 with somatically acquired *RAD51C* dysfunction; the magnitude of this finding suggests that *RAD51C* methylation is an important determinant of basal-like breast tumors. Preclinical models that lack *RAD51C* have been reported to exhibit sensitivity to PARP inhibitors[46], raising potential translational implications for this finding. Therefore, *RAD51C* methylation could aid in patient stratification for agents that target HRD, as is currently being done for ovarian cancer patients[47].

In contrast to *BRCA1/2*, *RAD51C*, and *PALB2*, protein-truncating germline variants in *ATM*, *CHEK2*, and *NBN* that were shown to confer increased risk of developing breast cancer[15] (including established pathogenic variants; **Supplementary Table 1**) were not associated with signature 3. This finding suggests that these events may drive cancer risk via mechanisms independent of the BRCA1/2 DSB-repair pathway**.** Indeed, in two pancreatic cancer studies with 17 combined cases of *ATM* mutations, 16 of which were biallelic, only 1 case showed a signature consistent with HRD[26,27]. In contrast, alterations among HR components that are in the same protein complex with BRCA1/2 (e.g., RAD51C, PALB2, and BARD1) are associated with signature 3, accounting for 8% of samples in the top quartile of signature 3 activity and 13% of cases without *BRCA1/2* events (**Fig. 4c**). These findings may further extend the rational use of novel therapeutics.

"BRCAness" refers to HRD in a tumor that lacks a germline *BRCA1* or *BRCA2* deleterious variant[16]. As we have shown, signature 3 can serve as a novel readout of HRD, and thus contribute to BRCAness, probably reflecting other defects in the *BRCA1/2* part of the pathway. It will be important to test the association among different measures of BRCAness, including ones based on expression, mutational signatures, and other molecular measures, as well as responses to treatment in clinical trials. Our measure of signature 3 could represent a putative biomarker for HR status. Current predictors of PARP-inhibitor or cisplatin sensitivity have had notable clinical success in a limited number of patients[48–50], which indicates the need for additional biomarkers such as HRD[48] or signature 3 to identify other individuals who could also benefit from these treatments.

The development of a companion diagnostic based on this work could inform the appropriate allocation of HRD-based therapies. Genomic-based scores, especially when based on clonal mutations, reflect the aggregated effect of HRD throughout the development of the cancer, but not necessarily its current HRD status[51]. Although not stated by the authors of ref. 51, ovarian cancers with *BRCA1* reversion[51] are likely to have retained high levels of signature 3 inherited from their ancestor cells, but are unlikely to have ongoing HRD. We expect a similar phenomenon with promoter methylation whereby platinum (or PARP-inhibition) resistance might arise as a result of demethylation of the relevant promoter (for example, *BRCA1* or *RAD51C*).

Our findings have notable implications for the analysis of germline variants. When a pathogenic *BRCA1/2* germline variant is found in the setting of a breast or ovarian cancer, it is inferred to be the putative driver of carcinogenesis[52]. However, among 28 carriers of pathogenic germline *BRCA1/2* variants in our series whose tumor LOH status was discernible, we identified four tumors that exhibited neither loss of the intact *BRCA1/2* allele nor elevated signature 3 activity. In these cases, either the tumors arose independently of BRCA1/2 inactivation, or *BRCA1/2* haploinsufficiency triggered events that, although cancer-initiating, were insufficient to independently generate HRD. *BRCA1* haploinsufficiency in breast epithelium can lead to defective

repair of stalled replication forks[53]. Inactivation of *BRCA2* among other HR factors may cause similar stalling. As these forks can be repaired by non-HR-dependent mechanisms[54], *BRCA1/2* haploinsufficiency could lead to non-HRD tumor phenotypes. Consequently, integration of signature 3 with *BRCA1/2* germline testing may affect treatment selection.

Moreover, the classification of missense germline variants remains a challenging prospect, given the breadth of uncommon variants and the large sample size often needed to reliably determine pathogenicity. Tumor signatures could be a powerful adjunct to existing methods for determining therapeutic approaches[55]. Importantly, our findings must be interpreted in the context of the TCGA study design. Namely, the samples from the TCGA may have limited generalizability to underrepresented minorities from whom few samples were collected. Similarly, accessibility to TCGA-participating centers and unblinded enrollment in this study may have introduced unforeseen biases into the study cohort. The clinical utility of signature 3 will require prospective validation, given the scarcity of pathogenic missense variants in this cohort ($n = 5$).

Although clinical genetics has long relied on the use of population-level data and family-based segregation analysis to identify high-risk patients, this approach is of somewhat limited utility among those who are part of small pedigrees or who otherwise have limited access to familial information, and in the context of the discovery of increasing numbers of very rare, possibly unique, variants. In the near future, with the increasing amounts of tumor-sequencing data becoming available[56], mutation signatures extracted from these clinical sequencing projects can be used to contribute to variant classification. It is possible that some VUSs could have a signature 3 profile that is sufficiently robust to support classification as likely benign (class 2) or likely pathogenic (class 4), even in the absence of other contributing data. In addition, the contribution of deleterious alterations in other genes, such as *PALB2* and *RAD51C*, to HR defects and hence to signature 3 could lead to the design of therapeutic trials based on the findings described herein.

METHODS

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

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

ACKNOWLEDGMENTS

L.Z.B. was supported by the Louis B. Mayer Foundation. L.W.E. was supported by grants from the Avon Breast Cancer Crusade and the Breast Cancer Research Foundation (BCRF). G.G. and J.K. were partially funded by the NIH TCGA Genome Data Analysis Center (U24CA143845). P.P., N.J.H., Y.E.M., and A.K. were funded by the startup funds of G.G. at Massachusetts General Hospital. A.D.D. was supported by grants from the Ludwig Center at Harvard and the Breast Cancer Research Foundation (BCRF). W.D.F. was supported by Susan G. Komen. G.G. was partly funded by the Paul C. Zamecnick, MD, Chair in Oncology at Massachusetts General Hospital.

AUTHOR CONTRIBUTIONS

P.P., J.K., and L.Z.B. conceived the work, performed analyses, and wrote the manuscript. R.K. and N.J.H. performed analyses and wrote the manuscript. G.T., D.R. and D.L. performed analyses. K.K. edited the manuscript. A.K., Y.E.M., and I.L. performed analyses and edited the manuscript. E.S.L. T.R.G., and A.Z. edited the manuscript. K.W.M. and A.O. wrote the manuscript. M.S.L. performed analysis and edited the manuscript. R.N.B. and C.C. provided data. D.A.H., L.W.E., and S.J.C. contributed scientific insight and edited the manuscript. P.W.L. and H.S. performed analysis and wrote the manuscript. A.D.D'A. conceived the work, contributed scientific insight, and edited the manuscript. W.D.F. and G.G. conceived the work, oversaw the analyses, and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

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

1. 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).
2. Antoniou, A. *et al.* Average risks of breast and ovarian cancer associated with *BRCA1* or *BRCA2* mutations detected in case series unselected for family history: a combined analysis of 22 studies. *Am. J. Hum. Genet.* **72**, 1117–1130 (2003).
3. Ceccaldi, R., Rondinelli, B. & D'Andrea, A.D. Repair pathway choices and consequences at the double-strand break. *Trends Cell Biol.* **26**, 52–64 (2016).
4. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
5. Burstein, M.D. *et al.* Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clin. Cancer Res.* **21**, 1688–1698 (2015).
6. Ellis, M.J. & Perou, C.M. The genomic landscape of breast cancer as a therapeutic roadmap. *Cancer Discov.* **3**, 27–34 (2013).
7. Ciriello, G. *et al.* Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* **163**, 506–519 (2015).
8. Ceccaldi, R. *et al.* Homologous-recombination-deficient tumours are dependent on Polθ-mediated repair. *Nature* **518**, 258–262 (2015).
9. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
10. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
11. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
12. Kim, J. *et al.* Somatic *ERCC2* mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016).
13. Kasar, S. *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, 8866 (2015).
14. Nielsen, F.C., van Overeem Hansen, T. & Sørensen, C.S. Hereditary breast and ovarian cancer: new genes in confined pathways. *Nat. Rev. Cancer* **16**, 599–612 (2016).
15. Easton, D.F. *et al.* Gene-panel sequencing and the prediction of breast-cancer risk. *N. Engl. J. Med.* **372**, 2243–2257 (2015).
16. Lord, C.J. & Ashworth, A. BRCAness revisited. *Nat. Rev. Cancer* **16**, 110–120 (2016).
17. Eccles, D.M. *et al. BRCA1* and *BRCA2* genetic testing—pitfalls and recommendations for managing variants of uncertain clinical significance. *Ann. Oncol.* **26**, 2057–2065 (2015).
18. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
19. Landrum, M.J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).
20. Meeks, H.D. *et al. BRCA2* polymorphic stop codon K3326X and the risk of breast, prostate, and ovarian cancers. *J. Natl. Cancer Inst.* **108**, djv315 (2015).
21. Scully, R. & Livingston, D.M. In search of the tumour-suppressor functions of BRCA1 and BRCA2. *Nature* **408**, 429–432 (2000).
22. Merajver, S.D. *et al.* Germline *BRCA1* mutations and loss of the wild-type allele in tumors from families with early onset breast and ovarian cancer. *Clin. Cancer Res.* **1**, 539–544 (1995).
23. Cornelis, R.S. *et al.* High allele loss rates at 17q12-q21 in breast and ovarian tumors from *BRCA1*-linked families. *Genes Chromosom. Cancer* **13**, 203–210 (1995).
24. Thompson, E.R. *et al.* Panel testing for familial breast cancer: calibrating the tension between research and clinical care. *J. Clin. Oncol.* **34**, 1455–1459 (2016).
25. Southey, M.C. *et al. PALB2*, *CHEK2* and *ATM* rare variants and cancer risk: data from COGS. *J. Med. Genet.* **53**, 800–811 (2016).
26. Waddell, N. *et al.* Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495–501 (2015).
27. Connor, A.A. *et al.* Association of distinct mutational signatures with correlates of increased immune activity in pancreatic ductal adenocarcinoma. *JAMA Oncol.* **3**, 774–783 (2017).
28. Livingston, D.M. Cancer. Complicated supercomplexes. *Science* **324**, 602–603 (2009).
29. Loveday, C. *et al.* Germline *RAD51C* mutations confer susceptibility to ovarian cancer. *Nat. Genet.* **44**, 475–476 (2012).
30. 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. *J. Clin. Oncol.* **33**, 304–311 (2015).

31. Blanco, A. *et al. RAD51C* germline mutations found in Spanish site-specific breast cancer and breast-ovarian cancer families. *Breast Cancer Res. Treat.* **147**, 133–143 (2014).
32. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
33. Pereira, B. *et al.* The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.* **7**, 11479 2016).
34. Foulkes, W.D. *et al.* Germline *BRCA1* mutations and a basal epithelial phenotype in breast cancer. *J. Natl. Cancer Inst.* **95**, 1482–1485 (2003).
35. Esteller, M. *et al.* Promoter hypermethylation and *BRCA1* inactivation in sporadic breast and ovarian tumors. *J. Natl. Cancer Inst.* **92**, 564–569 (2000).
36. Foulkes, W.D., Smith, I.E. & Reis-Filho, J.S. Triple-negative breast cancer. *N. Engl. J. Med.* **363**, 1938–1948 (2010).
37. Dietze, E.C., Sistrunk, C., Miranda-Carboni, G., O'Regan, R. & Seewaldt, V.L. Triple-negative breast cancer in African-American women: disparities versus biology. *Nat. Rev. Cancer* **15**, 248–254 (2015).
38. Rehm, H.L. *et al.* ClinGen—the Clinical Genome Resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
39. Spurdle, A.B. *et al.* ENIGMA—evidence-based network for the interpretation of germline mutant alleles: an international initiative to evaluate risk and clinical significance associated with sequence variation in *BRCA1* and *BRCA2* genes. *Hum. Mutat.* **33**, 2–7 (2012).
40. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
41. Hashizume, R. *et al.* The RING heterodimer BRCA1-BARD1 is a ubiquitin ligase inactivated by a breast cancer-derived mutation. *J. Biol. Chem.* **276**, 14537–14540 (2001).
42. Górski, B. *et al.* Founder mutations in the *BRCA1* gene in Polish families with breast-ovarian cancer. *Am. J. Hum. Genet.* **66**, 1963–1968 (2000).
43. Abkevich, V. *et al.* Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br. J. Cancer* **107**, 1776–1782 (2012).
44. Popova, T. *et al.* Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with *BRCA1/2* inactivation. *Cancer Res.* **72**, 5454–5462 (2012).
45. Youden, W.J. Index for rating diagnostic tests. *Cancer* **3**, 32–35 (1950).
46. Min, A. *et al.* RAD51C-deficient cancer cells are highly sensitive to the PARP inhibitor olaparib. *Mol. Cancer Ther.* **12**, 865–877 (2013).
47. Evans, T. & Matulonis, U. PARP inhibitors in ovarian cancer: evidence, experience and clinical potential. *Ther. Adv. Med. Oncol.* **9**, 253–267 (2017).
48. Chan, S.L. & Mok, T. PARP inhibition in *BRCA*-mutated breast and ovarian cancers. *Lancet* **376**, 211–213 (2010).
49. Fong, P.C. *et al.* Inhibition of poly(ADP-ribose) polymerase in tumors from *BRCA* mutation carriers. *N. Engl. J. Med.* **361**, 123–134 (2009).
50. Kaufman, B. *et al.* Olaparib monotherapy in patients with advanced cancer and a germline *BRCA1/2* mutation. *J. Clin. Oncol.* **33**, 244–250 (2015).
51. Patch, A.M. *et al.* Whole-genome characterization of chemoresistant ovarian cancer. *Nature* **521**, 489–494 (2015).
52. Clamp, A. & Jayson, G. PARP inhibitors in *BRCA* mutation-associated ovarian cancer. *Lancet Oncol.* **16**, 10–12 (2015).
53. Pathania, S. *et al. BRCA1* haploinsufficiency for replication stress suppression in primary cells. *Nat. Commun.* **5**, 5496 (2014).
54. Yeeles, J.T., Poli, J., Marians, K.J. & Pasero, P. Rescuing stalled or damaged replication forks. *Cold Spring Harb. Perspect. Biol.* **5**, a012815 (2013).
55. Lindor, N.M. *et al.* A review of a multifactorial probability-based model for classification of *BRCA1* and *BRCA2* variants of uncertain significance (VUS). *Hum. Mutat.* **33**, 8–21 (2012).
56. Hyman, D.M. *et al.* Precision medicine at Memorial Sloan Kettering Cancer Center: clinical next-generation sequencing enabling next-generation targeted therapy trials. *Drug Discov. Today* **20**, 1422–1428 (2015).

ONLINE METHODS

Data sets. We downloaded a total of 995 unique tumor–normal pairs for TCGA breast cancer BAM files from CGHub (https://cghub.ucsc.edu/; available at NCI's Genomic Data Commons (https://gdc.cancer.gov/)) on 2 June 2016 and applied the standard sample quality control and mutation-calling pipelines at Broad: MuTect[57] for SNVs, and Strelka[58] for indels. Application of downstream mutation quality control with the oxoG filter[59] and the panel of normal filtration identified 119,632 SNVs and 12,501 indels.

Mutation signature analysis. *Methods and algorithms.* The mutational-signature discovery process involves deconvoluting cancer somatic mutations counts, stratified by mutation contexts or biologically meaningful subgroups, into a set of characteristic patterns (signatures) and inferring the activity of each of the discovered signatures across samples. For this purpose we developed SignatureAnalyzer, which uses a Bayesian variant of NMF and was recently applied to several cancer genome projects (see ref.12 for additional background and technical details). The Bayesian NMF method exploits a 'shrinkage', or automatic relevance determination technique, to allow a sparse representation for both signatures and activities, as well as an optimal inference for the number of signatures (K), by iteratively pruning away irrelevant components. We used the same parameters described in ref.12. All SNVs were classified into 96 possible mutation types (or categories) on the basis of six base substitutions (C>A, C>G, C>T, T>A, T>C, and T>G) within the trinucleotide sequence context including the bases immediately 5′ and 3′ to the mutated base. We considered only SNVs on autosomal chromosomes (110,704 SNVs).

A subset of 750 (of the 995) cases was also recently analyzed by Alexandrov *et al.*[10], who used a slightly different methodology. In their analysis (which also included additional non-TCGA cases), they identified five mutational signatures: signature 1B (aging), signatures 2 and 13 (APOBEC), signature 3 (BRCA), and signature 8 (ref. 10). The results from their analysis are largely consistent with ours; however, to take full advantage of our Bayesian framework and to accurately represent the signatures found in our cohort, we performed *de novo* signature discovery and assignment of per-patient signature activities.

Signatures in 995 TCGA breast carcinoma samples (TCGA995). We first carried out a signature analysis for the 995 invasive breast carcinoma samples from TCGA (TCGA995)[4], and identified seven mutational signatures (**Supplementary Fig. 1**). Among them, four signatures were recurrently active across samples: S995A (cosine similarity 0.96 to COSMIC1), S995B (cosine similarity 0.89 to COSMIC3), S995C (a combined APOBEC signature of COSMIC2 and COSMIC13), and S995D (cosine similarity 0.83 to COSMIC6). A majority of SNVs (88%) were attributed to these four signatures—S995A (18%), S995B (26%), S995C (33%), and S995D (11%). The activities of the three remaining signatures—S995E (cosine similarity 0.96 to COSMIC10), S995F (cosine similarity 0.9 to COSMIC24), and S995G (no match to 30 COSMIC signatures with cosine similarity > 0.83)—were predominant each in a singular case: an ultra-mutant tumor (>5,800 SNVs) harbored a POLE exonuclease domain mutation (P286R) and contained about 70% of SNVs associated with S995E; 48% of SNVs associated with S995F were attributed to a single sample (and only 2% to the next highest active sample); and 88% of SNVs associated with S995G were attributed to a single sample (and only 4% to the next highest active sample).

Signatures in 992 TCGA breast carcinoma samples (TCGA992). We excluded three samples with significant enrichment of singleton signatures (S995E, S995F, and S995G) and carried out the signature analysis again across the 992 samples (TCGA992). Out of 50 independent Bayes NMF runs, nearly 80% of the runs converged to the four-signature solution (TCGA992-S4; **Fig. 1**), and 20% converged to the five-signature solution (TCGA992-S5; **Supplementary Fig. 16**). Although signatures in both solutions looked similar, there were some notable differences: (i) both C>T APOBEC and C>G APOBEC signatures in TCGA992-S5 merged to form a single combined APOBEC signature in TCGA992-S4 (APOBEC in **Fig. 1**), and (ii) the concordance of putative BRCA and MSI signatures to corresponding COSMIC signatures was much higher in the four-signature solution with cosine similarities of 0.95 versus 0.82 to COMSIC3 and 0.95 versus 0.82 to COMSIC3, in TCGA992-S4 and TCGA992-S5, respectively. For downstream analyses throughout, we used the signature set in TCGA992-S4, which was the most probable (i.e., had a higher posterior probability) and had a much higher concordance to COSMIC signatures.

Germline variant calling, sample QC, and site QC. The variants in the 992 germline exomes of the TCGA breast cancer study were jointly called with about 36,600 additional germline exomes (a total of 37,607 germline exomes,) including TCGA samples, non-TCGA cancer samples, and general population controls. The variants were called using best practices with the Genome Analysis Toolkit HaplotypeCaller (version 3.1)[60].

Germline variants among 20 DNA-repair genes commonly used in clinical gene panel tests (*BRCA1*, *BRCA2*, *RAD51C*, *PALB2*, *ATM*, *ATR*, *BARD1*, *BRIP1*, *MRE11*, *NBN*, *PTEN*, *RAD50*, *RAD51D*, *TP53*, *XRCC2*, *FANCA*, *FANCD2*, *FANCM*, *CHEK1*, and *CHEK2*) were extracted from the call set, and common variants (with a minor allele frequency > 1% in the non-cancer ExAC[61] normal population cohort (ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/subsets/)) were removed. All genotype calls with a genotype quality score less than 20 (the Phred-scaled confidence in the genotype call) were removed, along with sites that violated Hardy–Weinberg equilibrium ($P < 0.001$) and sites with a quality-by-depth score of <5.

Clinical annotation of *BRCA1/2* germline variants. *BRCA1/2* germline events were considered to be pathogenic if they resulted in a premature stop codon (i.e., nonsense mutation) or a frameshift (excluding *BRCA2* p.Lys3326*, which is known to be a low-penetrance allele and is therefore a special case). The majority of these mutations are represented in ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/). In addition, missense mutations that were annotated as pathogenic in ClinVar were also considered pathogenic in our analysis. We treated splice-site variants similarly to missense mutations. However, no pathogenic splice-site variants were found in *BRCA1/2* in this cohort. In addition, for each of the heterozygous germline variants, we also annotated their allelic imbalance in the corresponding tumor. We assessed the allelic imbalance by tallying reads that supported the reference and the nonreference alleles at the germline site in the corresponding tumor whole-exome sequencing data (BAM file). To determine whether the alternate (ALT) allele was lost in cases of LOH, we carried out a one-sided Fisher's exact test of the reads that supported the ALT/reference (REF) allele in the tumor versus the patient-matched normal tissue to establish whether the bias toward the ALT increased in the tumor. Cases in which the test yielded $P < 0.05$ were considered 'ALT retained'. For $P > 0.05$, we defined the LOH status as undetermined.

A file with a table of germline variants, their annotations, and their associated samples is available in a TCGA patient-protected manner in FireCloud (available from the corresponding author upon request).

Loss-of-heterozygosity analysis. To determine the local copy numbers of *BRCA1* and *BRCA2*, we ran the ABSOLUTE algorithm[62] on all TCGA breast cancer tumor samples, which yielded an allele-specific absolute copy number profile for the entire genome. The ABSOLUTE algorithm was run on both exome-based and SNP-array-based copy-ratio segmentation, and the higher-quality result of the two was used in the final copy-number analysis. Absolute copy numbers at the *BRCA1/2* loci were estimated from the genome-wide profile. A gene was labeled as having LOH when one allele was clonally deleted (i.e., had an allelic absolute copy number of 0) and the other allele was not clonally deleted (i.e., had an absolute copy number > 0). Homozygous deletions occurred when the allelic copy number was 0 for both alleles. This analysis was also conducted with FireCloud (http://www.firecloud.org), a cloud-based genome-analysis platform developed at the Broad Institute.

Epigenetic silencing. DNA methylation of all gene promoters was assessed at probes within a window that spanned 1,500 bp upstream and downstream of transcription start sites, using both HM27 and HM450 data in 975 of the 992 samples. To evaluate whether methylation of a promoter was associated with transcriptional downregulation (i.e., epigenetic silencing), we correlated promoter-methylation levels with mRNA levels. Among HR genes, the promoter-methylation levels of *BRCA1* and *RAD51C* were significantly

anticorrelated with the gene mRNA levels ($P < 0.001$; Wilcoxon rank-sum test) and, therefore, were identified as 'epigenetically silenced'. We further evaluated epigenetic silencing on a per-sample basis, using promoter-methylation levels independent of concomitant gene expression levels.

CoMut plot data. The coMut plot in **Figure 2** is a presentation of sample-level information that is found in a summary table in a TCGA patient-protected manner in FireCloud and is available from the corresponding author upon request.

Data availability. All the raw data that were used in this analysis are part of the TCGA breast cancer project and are available at the NCI's Genomic Data Commons (https://gdc.cancer.gov/). A **Life Sciences Reporting Summary** for this paper is available.

57. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol* **31**, 213–219 (2013).
58. Saunders, C.T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
59. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41**, e67 (2013).
60. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
61. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
62. Carter, S.L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).

Corresponding Author: Gad Getz

Date: Jun 8, 2017

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined.

   Sample sizes were determined by the number of patients in each of the published cohorts. See "Data sets" section of Methods for additional details regarding cohorts.

2. Data exclusions

   Describe any data exclusions.

   N/A

3. Replication

   Describe whether the experimental findings were reliably reproduced.

   N/A

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   N/A

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   N/A

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	☐	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *p* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Nature Genetics: doi:10.1038/ng.3934

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	The source code for SignatureAnalyzer is available at https://www.broadinstitute.org/cancer/cga

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	The genomic data is available at NCI's Genomic Data Commons (https://gdc.cancer.gov/)

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	N/A
b. Describe the method of cell line authentication used.	N/A
c. Report whether the cell lines were tested for mycoplasma contamination.	N/A
d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	N/A

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	N/A

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Participants were breast cancer patients as part of the TCGA project

Nature Genetics: doi:10.1038/ng.3934