nature genetics

TECHNICAL REPORT

https://doi.org/10.1038/s41588-019-0390-2

Detecting the mutational signature of homologous recombination deficiency in clinical samples

Doga C. Gulhan[1], Jake June-Koo Lee[1], Giorgio E. M. Melloni[1], Isidro Cortés-Ciriano[1,2] and Peter J. Park[1*]

Mutations in *BRCA1* and/or *BRCA2* (*BRCA1/2*) are the most common indication of deficiency in the homologous recombination (HR) DNA repair pathway. However, recent genome-wide analyses have shown that the same pattern of mutations found in *BRCA1/2*-mutant tumors is also present in several other tumors. Here, we present a new computational tool called Signature Multivariate Analysis (SigMA), which can be used to accurately detect the mutational signature associated with HR deficiency from targeted gene panels. Whereas previous methods require whole-genome or whole-exome data, our method detects the HR-deficiency signature even from low mutation counts, by using a likelihood-based measure combined with machine-learning techniques. Cell lines that we identify as HR deficient show a significant response to poly (ADP-ribose) polymerase (PARP) inhibitors; patients with ovarian cancer whom we found to be HR deficient show a significantly longer overall survival with platinum regimens. By enabling panel-based identification of mutational signatures, our method substantially increases the number of patients that may be considered for treatments targeting HR deficiency.

Mutational signature analysis has emerged as a powerful approach for investigating the processes that generate somatic mutations. Conceptually, this analysis is based on the observation that different mutational processes generate specific base-pair changes, typically in particular nucleotide contexts[1]. For instance, ultraviolet radiation generally results in C-to-T changes, often with a C flanked by a C or T on the 5′ side. In its popular form[2,3], this analysis computes a vector of 96 triplets (six substitution subtypes, C>A, C>G, C>T, T>A, T>C and T>G, each flanked by one of the four types on the 5′ and 3′ sides) for a set of genomes and deconvolves the observed mutational spectra into independent components. Application of this concept to thousands of tumor samples with exome or whole-genome sequencing (WGS) has led to a catalog of nearly 40 mutational signatures operative in cancer[2,4]. Some of these signatures have been matched to specific mutational processes, both endogenous (for example, replication clock, apolipoprotein B mRNA editing catalytic polypeptide-like (APOBEC) cytosine deaminases, defects in the DNA repair machinery) and exogenous (for example, smoking carcinogens, ultraviolet radiation)[5–12], although the majority still remain uncharacterized.

In breast cancer, a landmark study of 560 whole genomes[13] and subsequent studies[14,15] showed that one of these signatures—'Signature 3' (Sig3)—corresponds to a deficiency in the HR machinery (Supplementary Fig. 1). This signature is observed in tumors with complete *BRCA1/2* inactivation, which can occur by germline and somatic point mutations combined with loss of heterozygosity, hypermethylation of *BRCA1* promoters or loss-of-function mutations of *PALB2* and *RAD51D*[15]. Experimentally, Sig3 was observed in *BRCA*$^{-/-}$ isogenic cell lines, providing direct evidence of its association with HR deficiency[16].

Importantly, there is increasing evidence that Sig3 is not limited to those with a germline mutation in *BRCA1*/2 or other known HR-related genes[13,15,17]. This is clinically relevant because those without a mutation in a known HR gene but who present Sig3 may benefit from treatments that target selective vulnerabilities of HR-deficient cancers. A recent study using breast cancer organoids, for example, has shown that a high burden of Sig3 mutations is associated with a better response to poly(ADP-ribose) polymerase (PARP) inhibitors[18]. Inhibitors of PARP enzymes cause multiple double-strand breaks; tumor cells that cannot repair the breaks because of HR deficiency do not survive.

In this study, we propose a new method for detecting Sig3 from sequencing data of an individual. Although previous methods have addressed the identification of HR deficiency through mutational signatures[14,15], they were limited to exome or whole-genome data, thus hampering use in clinical practice. For the most common genetic testing platform in oncology clinics—targeted sequencing panels—the number of mutations identifiable is far too small for standard signature analysis. A recent panel-based study of 10,000 cancer patients, for example, could perform signature analysis for only 6% of the samples with the highest mutational burden[19]. Our computational tool, Signature Multivariate Analysis (SigMA), uses a likelihood-based approach that can detect signatures, including Sig3, from low mutation counts. Thus, application of this method has the potential to vastly expand the number of patients that could benefit from treatments available for HR-deficient tumors.

Results

Limitations of current methods. Existing methods for signature analysis follow one of two approaches. One approach is to discover signatures from all available genomes by applying an unguided decomposition algorithm, such as non-negative matrix factorization (NMF)[3,20,21]. The other approach is to find an optimal combination of predefined signatures for a given sample, for example, by using non-negative least squares (NNLS)[20,22]. The commonality in the two approaches is the decomposition step where the mutational spectra of tumors are described as a linear combination of signatures. In the first case, the signatures are discovered simultaneously with their coefficients, which we also refer to as 'exposures'; in the

[1]Department of Biomedical Informatics and Ludwig Center at Harvard, Harvard Medical School, Boston, MA, USA. [2]Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Cambridge, UK. *e-mail: peter_park@hms.harvard.edu

second case, a set of signatures is given, and the algorithm determines their exposures.

However, these methods are inadequate when the number of mutations is small. The NMF approach is unguided and therefore requires more information than the NNLS method. When there is insufficient information—that is, too few genomes or too few mutations per genome—only a subset of signatures that cause high mutational burden or are active in the vast majority of genomes are discovered, leading to low sensitivity (Supplementary Note; Supplementary Fig. 2). Moreover, the spectrum of a single signature is often affected by other signatures active in the same dataset; for example, signatures with correlated exposures may not be separated into distinct components (Supplementary Note for other computational issues). On the other hand, the second approach cannot be used for de novo signature discovery and requires the user to select the signatures to be included in the decomposition on the basis of prior knowledge. If it is not constrained, the NNLS-based method frequently leads to misidentification of signatures (low specificity) because the optimal solution may not be unique when there are many similar signatures in the catalog. This issue becomes more severe when few mutations are present (Supplementary Note; Supplementary Figs. 3 and 4).

The SigMA algorithm. SigMA enables accurate identification of mutational signatures even when the mutation count is very small. It combines the elements of the approaches explained earlier with new measures for associating mutations to signatures. First, it replaces the error-prone spectrum decomposition step with a clustering step, using the rich resource of existing WGS data that inform us of the co-occurring signatures and their relative contributions to a given tumor type and its subtypes. After identifying clusters of samples with similar mutational spectra for each tumor type (Methods and Supplementary Note), we compare the mutational spectrum of a new sample to each of the cluster averages by using the likelihood-based similarity measure described later. This allows the classification of the new tumor together with tumors that share similar combinations of signatures. When the mutation count is low, this is a more stable approach for inferring a combination of signatures present in a sample than performing a linear decomposition directly (Supplementary Note).

For breast cancers, we started with 730 WGS samples, of which 67 (9%) had biallelic inactivation of *BRCA1/2*, and obtained 12 clusters (Fig. 1a). These clusters fall broadly into four categories: (1) Sig3$^+$; (2) predominantly APOBEC[1,5,23]; (3) dominant 'Clock'[24]; and (4) microsatellite instability (MSI). On the basis of these clusters, we assigned a new sample to be, for example, Sig3$^+$ when the most similar cluster average was from the Sig3$^+$ category. The results of clustering in some other tumor types are shown in Fig. 1b and Supplementary Fig. 5a. The differences in the prevalence of signatures in each tumor type (for example, higher prevalence of Sig3 in ovarian cancer compared to pancreatic cancer) and the differences in the common signatures across tissues (for example, APOBEC signatures in bladder cancer and Signature 17 in stomach and esophageal cancer) necessitate a tumor type-specific procedure.

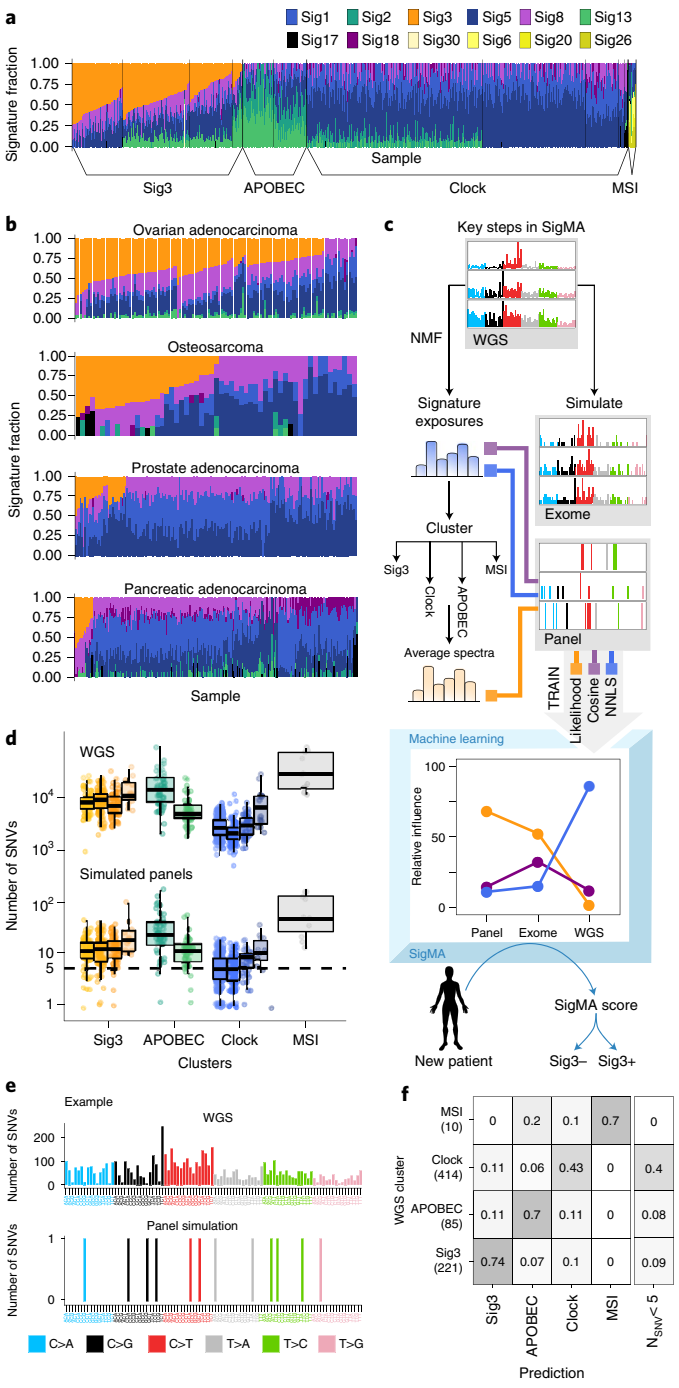

Fig. 1 | Overview of SigMA for Sig3 prediction. a, The 730 WGS breast cancers are clustered according to their fractional signature compositions; the resulting clusters are grouped into four categories (Sig3, APOBEC, Clock, MSI). **b**, Same clustering analysis for other tumor types (further detail in Supplementary Fig. 3). **c**, Key steps in SigMA. Signature analysis is performed with WGS data and the signature exposures of each sample are determined. Simulated exomes and panels are generated by subsampling from WGS data. For the simulated data, several statistics (for example, cosine similarity, likelihood) are calculated. A machine-learning classifier is trained using these features. For panels, the likelihood measure receives more weight in the prediction. **d**, Number of SNVs for the WGS samples ($n=780$) and simulated panels ($n=560$). Those panels with zero mutation ($n=170$) are not shown. There is a three-orders-of-magnitude reduction in the number of SNVs for panels compared to WGS. Each box represents a cluster ($n=66$, 87, 55, 13, 42, 43, 227, 134, 41, 12 and 10, respectively); three small Clock clusters are merged into one. The dashed horizontal line marks five mutations, the minimum required for inference. **e**, Example showing how sparse the observed spectrum is for a simulated panel compared to the WGS spectrum of the same sample. On the *x* axis, the 16 possible trinucleotide contexts are repeated for each of the six substitution types. **f**, Comparison of categories obtained by WGS analysis (*y* axis) and by SigMA (*x* axis) on simulated panels at a 10% FPR. The rows add up to 1 and the number of samples is given below the categories. The samples with fewer than five SNVs in the simulated panels are not classified (column furthermost to the right).

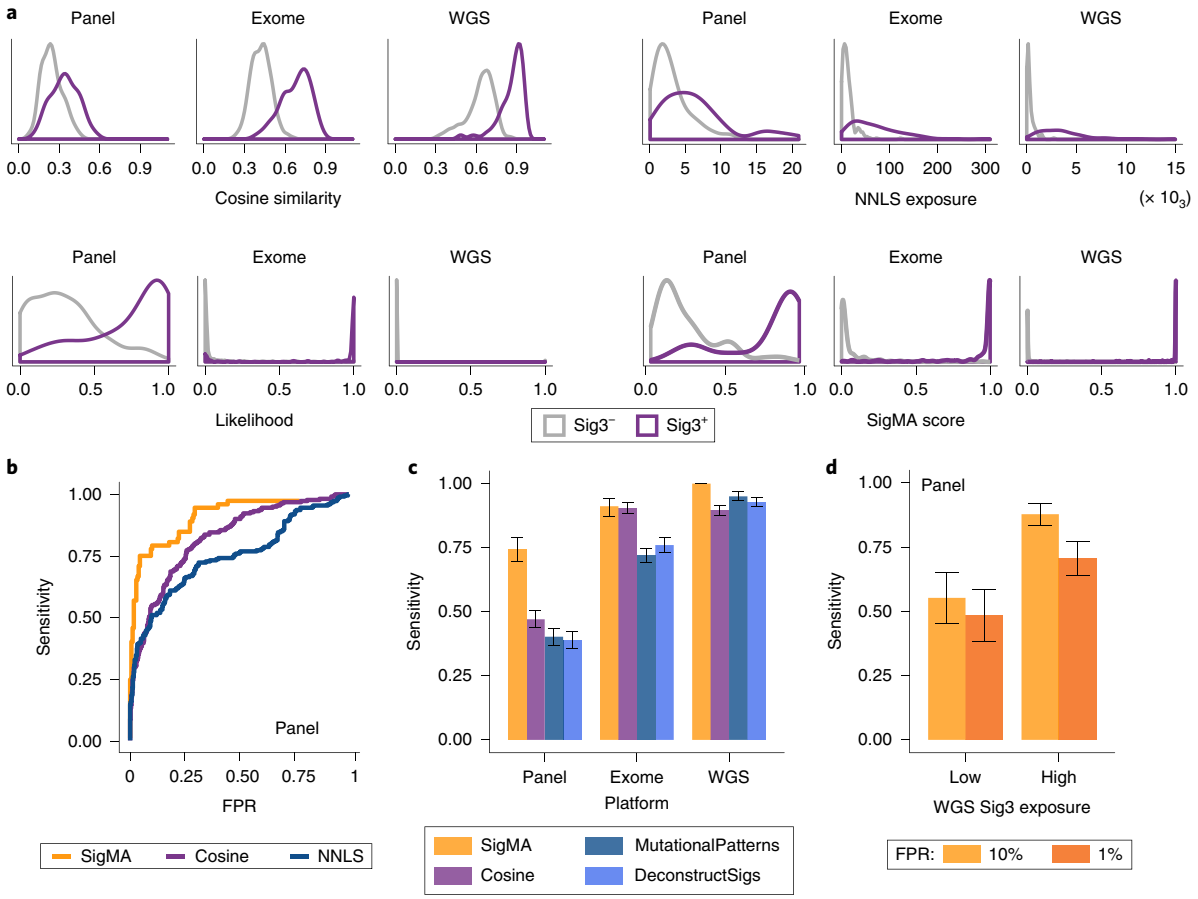

Fig. 2 | Performance of SigMA. a, The density distributions of four measures (cosine similarity, NNLS exposure, likelihood and SigMA score) are shown for Sig3$^+$ (purple) and Sig3$^-$ (gray) tumors and for the three platforms (panel, exome and WGS; the first two are simulated from the WGS; see Supplementary Note). Reference Sig3 status is determined by WGS-based NMF analysis (Supplementary Note). **b**, Sensitivity versus FPR for SigMA compared to the stand-alone use of cosine similarity and our implementation of NNLS (see Supplementary Fig. 10 for exomes and WGS). **c**, Sensitivity of SigMA for detecting Sig3 is compared to cosine similarity and two NNLS-based tools[20,22]. FPR was fixed at 10% for the panels and at 5–8% for exomes and WGS. **d**, Increased sensitivity when Sig3 exposure is high. The samples are divided into high/low exposure groups according to the median exposure. For panels **c** and **d**, the error bars denote the s.e.m.

The second component of SigMA is the similarity measure used for matching the mutational pattern of a given sample to the profiles of the clusters. A standard measure for comparing two spectra has been the cosine similarity[2], which is the cosine of the angle between two vectors in space. This measure is inadequate for identifying signatures when the mutation count is small because it is sensitive to minor changes in the mutational spectrum; even a single mutation can cause a large deviation in the angle. We propose a much more robust and statistically sound approach: we calculate the likelihood of the mutations in the new sample to have been generated from the probability distribution defined by the mutational profiles of each tumor cluster (Methods and Supplementary Note). A simple coin tossing example that illustrates the differences between the two methods can be found in Supplementary Fig. 6; similar demonstrations for mutational signatures are shown in Supplementary Figs. 7 and 8 and in the Supplementary Note.

To develop a unified framework that applies equally well to different types of sequencing platforms (panels, exomes and WGS), we combine several variables commonly used in signature analysis with our likelihood measure in a multivariate form (Fig. 1c). Thus, whether the most informative measure is the likelihood calculated from the average spectra (for panels) or linear decomposition accompanied by likelihood (for WGS), our method handles it automatically. The weighting of different components is done with gradient boosting classifiers (Supplementary Fig. 9d), which we found to be less sensitive to noise and more efficient compared to other machine-learning methods (Supplementary Figs. 10 and 11 and Supplementary Note).

Application of SigMA to simulated panel data. To illustrate the advantages of SigMA, we generated simulated datasets mimicking two widely used panels, MSK-IMPACT (Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets)[19] and FoundationOne[25]. The simulation is done by downsampling the 730 WGS samples, whose signature decomposition will serve as the gold standard (Supplementary Note). For a 410-gene panel covering a 1.7 Mb capture region (MSK-IMPACT), the number of mutations is typically reduced by about 1,000-fold (Fig. 1d), with the distribution of mutation counts similar to that observed for real data (see next section; Supplementary Fig. 12a). Among the 221 Sig3$^+$ samples, the average mutation count is 11.3, where 19 (8.6%) tumors have fewer than five mutations.

The sparsity of the simulated mutational spectrum for a Sig3$^+$ tumor in comparison to its WGS counterpart (Fig. 1e and Supplementary Fig. 13) illustrates the difficulty of making inferences about mutational signatures using panel-based data. Despite the large reduction in the mutation count, the SigMA classification of simulated panels for Sig3$^+$ cases mostly agree with

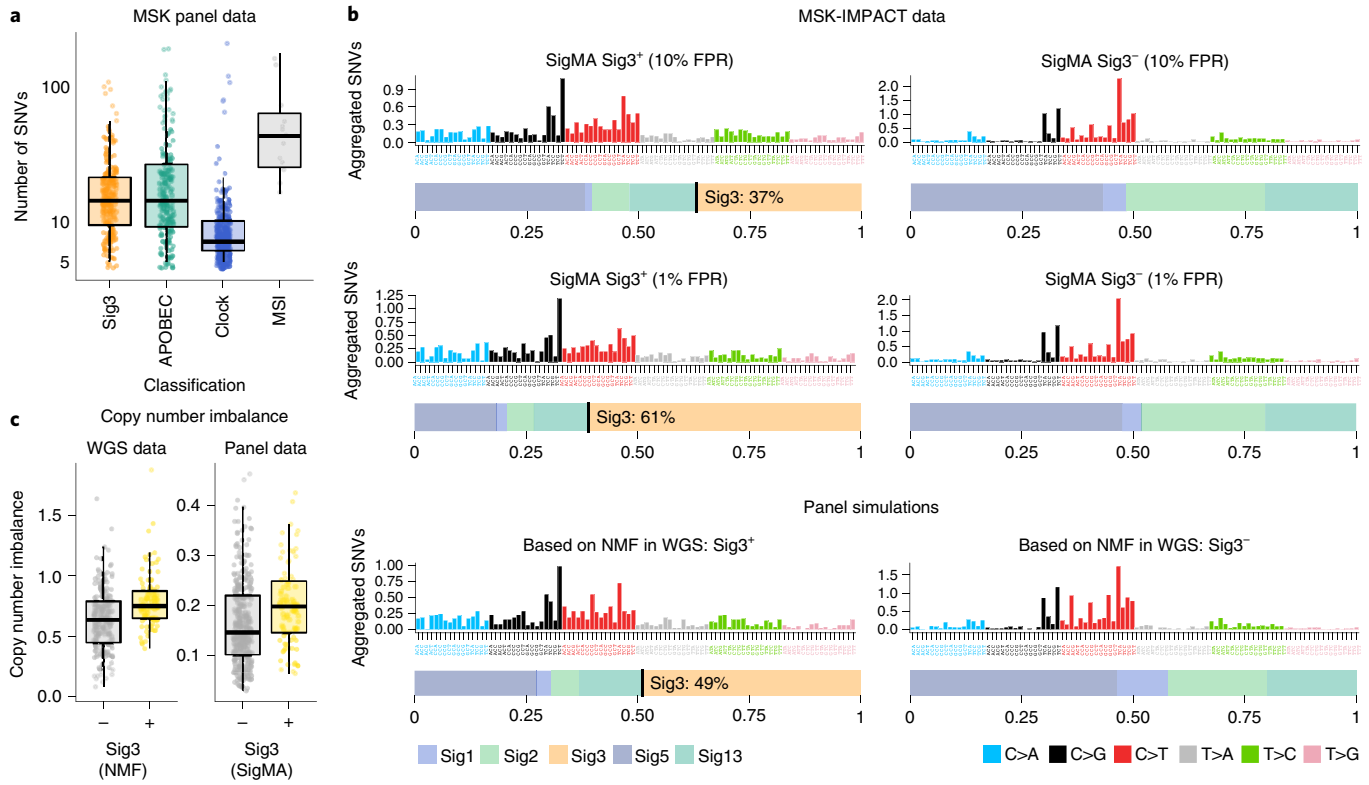

Fig. 3 | Validation of SigMA on MSK-IMPACT data. a, Total number of mutations in the MSK-IMPACT panel data by SigMA classifications. A large number of cases have five to ten mutations; the number of mutations in each predicted category ($n = 202, 176, 270, 13$) is similar to that of the simulated panels shown in Fig. 1d. Both high-confidence and low-confidence Sig3$^+$ samples are included in the Sig3$^+$ category. **b**, Average mutational spectra of tumors classified as Sig3$^+$ or Sig3$^-$ by SigMA. The first two rows correspond to modest (10% FPR) and stringent (1% FPR) criteria. These spectra resemble those from the simulated panels (third row), which are grouped on the basis of the WGS data. The horizontal bar below each spectrum shows the fractions of signatures found by decomposing the average spectra by NNLS. **c**, Among the WGS cases, Sig3$^+$ cases ($n = 221$) show a higher copy number imbalance than Sig3$^-$ cases ($n = 509$). The MSK panel samples split according to the SigMA classification ($n = 202$ and 459) show similar differences in copy number imbalance (Supplementary Note).

Table 1 | Sig3$^+$ cases identified in different tumor types by SigMA[a]

	Simulations			MSK-IMPACT data			Combined
Tumor type	**Samples**	**Sig3$^+$**	**%**	**Samples**	**Sig3$^+$**	**%**	**%**
Ovarian cancer	221	86	38.9	73	27	37.0	38.0
Osteosarcoma	47	12	25.5	31	9	29.0	27.3
Endometrial carcinoma	44	10	22.7	109	16	14.7	18.7
Breast cancer	731	138	18.9	878	121	13.8	16.3
Bladder cancer	23	3	13.0	322	56	17.4	15.2
Medulloblastoma[b]	123	16	13.0	-	-	-	13.0
Prostate adenocarcinoma	158	18	11.4	518	70	13.5	12.5
Ewing's sarcoma[b]	87	8	9.2	-	-	-	9.2
Pancreatic adenocarcinoma	230	14	6.1	361	29	8.0	7.1
Stomach adenocarcinoma	35	2	5.7	80	5	6.3	6.0
Pancreatic neuroendocrine cancer	62	3	4.8	54	3	5.6	5.2
Esophageal adenocarcinoma	296	17	5.7	71	3	4.2	5.0

[a]Strict threshold settings, with a 1–5% FPR, are used. This corresponds to a 2–7% FDR for tumor types with high prevalence of Sig3 and a 19–67% FDR for tumor types with low prevalence of Sig3. [b]These tumor types do not have a sufficient number of mutations in the panels. Instead the number of cases identified in exome simulations are shown.

the true categories defined from the WGS (163 out of 221 Sig3$^+$ tumors in the WGS analysis are correctly classified by SigMA; accuracy = 0.84 at a 10% false positive rate (FPR); Fig. 1f). Further classification of Sig3$^-$ tumors into Clock, APOBEC and MSI groups is 73% accurate for cases that have at least five mutations (Supplementary Note).

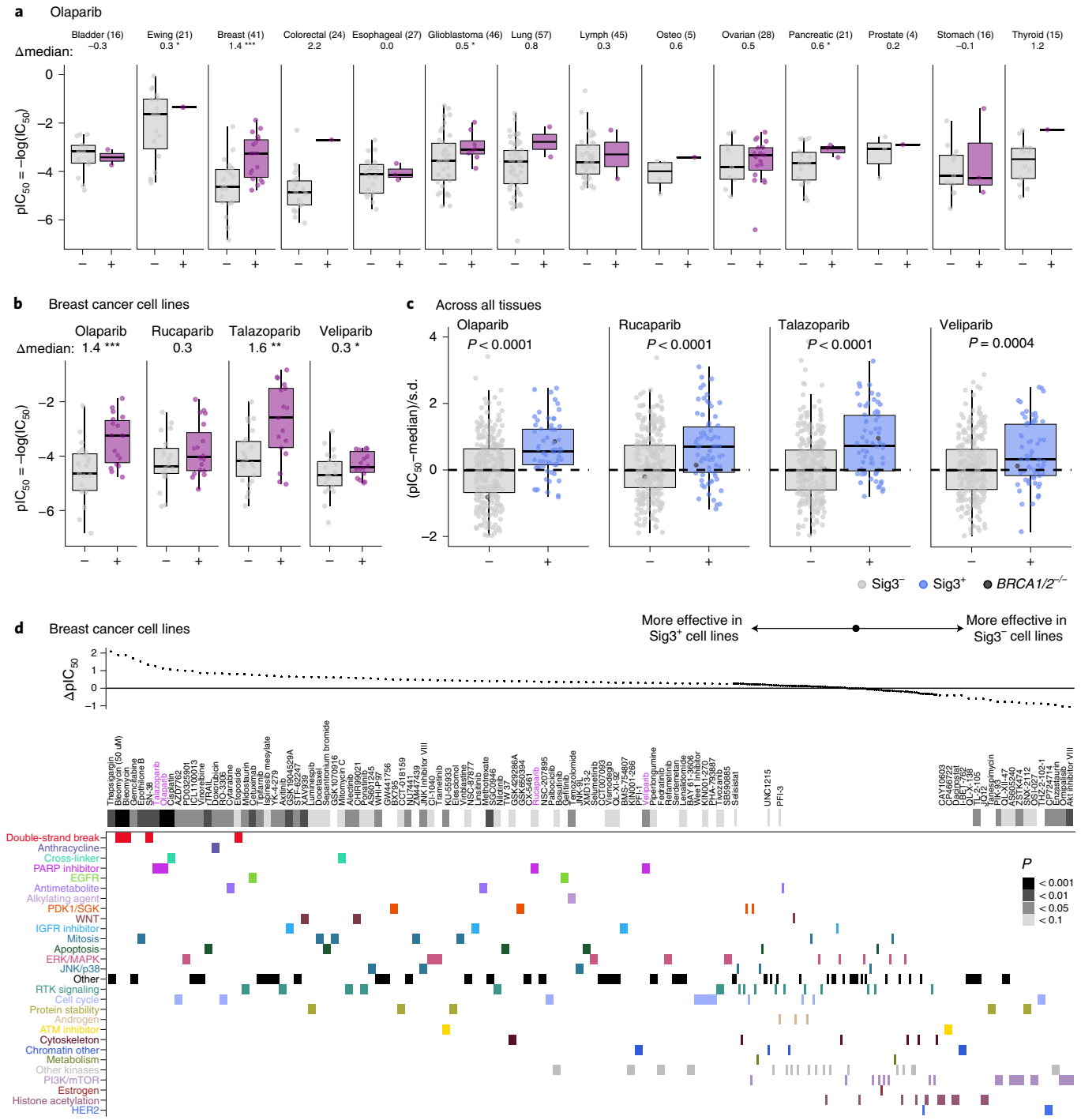

Fig. 4 | Experimental validation using drug response data. a, pIC_{50} values for olaparib in cell lines from different tumor types for $Sig3^+$ and $Sig3^-$ cell lines. The numbers of samples are in parentheses; the numbers below the tissue names are the differences in the median pIC_{50} values in $Sig3^+$ and $Sig3^-$ groups. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. **b**, The pIC_{50} values of olaparib, rucaparib, talazoparib and veliparib for breast cancer cell lines ($n = 41$, 39, 41 and 40) show significant differences for the two groups. **c**, Combined results for the four PARP inhibitors for the cell lines ($n = 366$, 371, 369 and 340) across all tumor types after normalization in each tumor type by subtracting the mean of the $Sig3^-$ cell lines and dividing by the s.d. For panels **a**–**c**, the *P* values are calculated using one-sided *t*-tests. **d**, Drug response difference in $Sig3^+$ and $Sig3^-$ breast cancer cell lines ($n = 36$–43) to PARP inhibitors across drug classes. The difference in the mean pIC_{50} values, denoted as ΔpIC_{50}, for the $Sig3^+$ and $Sig3^-$ tumors are shown at the top. The drugs with the largest difference are shown on the left. The *P* values from the one-sided *t*-tests are depicted in the heat map below the drug names. In the main panel, drugs are grouped according to their mechanism of action (*y* axis), which are ranked according to the position of the drugs in that category on the *x* axis. To enhance readability, the names of the drugs with small ΔpIC_{50} values are removed and these columns are contracted. ATM, serine-protein kinase ATM; EGFR, epidermal growth factor receptor; ERK, extracellular signal-regulated kinase; IGFR, insulin-like growth factor 1 receptor; JNK, c-Jun N-terminal kinase; MAPK, mitogen-activated protein kinase; mTOR, mammalian target of rapamycin; p38, p38 mitogen-activated protein kinase; PDK1, pyruvate dehydrogenase kinase isoform 1; PI3K, phosphoinositide 3-kinase; RTK, receptor tyrosine kinase; SGK, serine/threonine-protein kinase; WNT, Wnt signaling pathway.

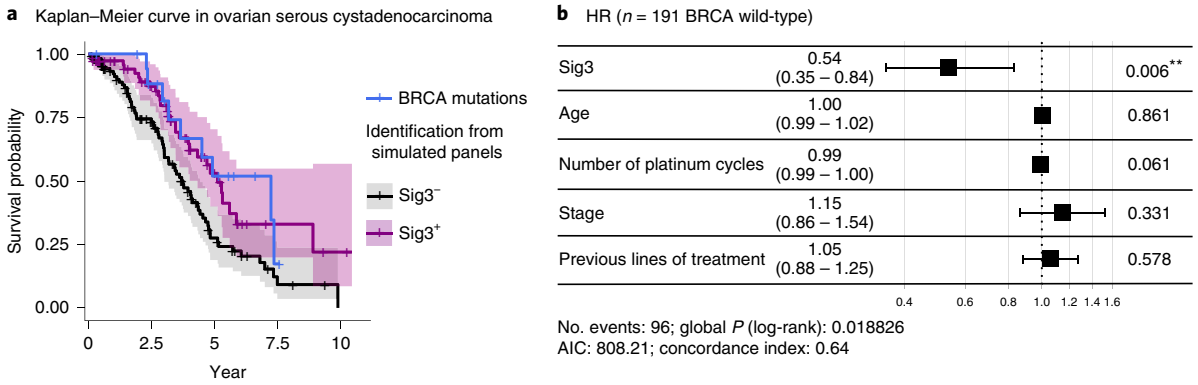

Fig. 5 | Survival analysis for patients with Sig3⁺ ovarian cancer. a, Kaplan–Meier curves for *BRCA1/2*-wild-type Sig3⁺ (pink; $n=76$) and Sig3⁻ (gray; $n=115$) patients. The patients with *BRCA* mutations (blue; $n=19$) are shown separately. Sig3 is identified from panels simulated from exome data. The shaded areas correspond to 95% CIs. Patients were treated with platinum regimens. **b**, Hazard ratios for the variables in the Cox regression model. The horizontal bars show the 95% CI; the *P* values are calculated with two-sided likelihood-ratio tests. Only the cases with at least five mutations in the simulated panels are considered in this analysis. AIC, Akaike information criterion.

In Fig. 2a, we show that cosine similarity and NNLS show reasonably good separation between the Sig3⁺ and Sig3⁻ cases for WGS data but not for panel data. In contrast, SigMA shows much better separation, especially for the panel data (Fig. 2a). The receiver operating characteristic (ROC) curves for panels (Fig. 2b) show that SigMA achieves higher sensitivity at the same FPR compared to other methods. The multivariate formulation of SigMA results in further improvement for all platforms (Supplementary Fig. 9a,e). At an FPR of 10% (25% false discovery rate (FDR)), SigMA has a sensitivity of 74% for panels, which is markedly higher than the 37–47% obtained for other methods (Fig. 2c). In another set of simulated panels with lower genomic coverage (FoundationOne with 315 genes; 253 genes in common with MSK-IMPACT), the sensitivity is slightly lower (68%) (Supplementary Fig. 9b). When Sig3 composes a large component of the mutational spectrum, sensitivity for detection tends to be substantially higher (Fig. 2d), which might be clinically relevant because Sig3⁺ tumors with a larger number of mutations belonging to Sig3 may be more responsive to PARP inhibitor treatment[18].

In WGS breast cancers, the number of predicted Sig3⁺ cases that do not have biallelic inactivation of *BRCA1/2* is 2.2-fold higher than those that do, indicating that a substantial number of cases that may benefit from treatments targeting HR deficiency is missed with the current *BRCA*-based criterion. For clinical use, we can tune the SigMA parameters to lower the FPR to 1% (4% FDR), at the cost of decreasing sensitivity to 50%. We refer to the tumors that pass the more stringent threshold as high-confidence Sig3⁺ cases; the remaining positive cases are referred to as low-confidence cases. The relative fraction of *BRCA1/2* mutant tumors among the high-confidence cases is still about 1 in 3. Thus, even with only high-confidence calls, SigMA will double the number of breast cancer patients determined to have HR deficiency compared to using *BRCA1/2* status alone.

Detection of Sig3 in the MSK-IMPACT panels. To validate the performance of SigMA on real panel data, we applied it to the 878 breast tumors profiled on the 410-gene MSK-IMPACT panel[19]. For tumors with at least five mutations, we classified them into the same four categories (Fig. 3a). We detected 202 cases (23%) that are likely to be Sig3⁺, with 121 (14%) passing a more stringent selection criterion.

When we aggregate all the mutations found in Sig3⁺ cases predicted by SigMA (Fig. 3b, top two rows) and compare their spectrum to that obtained from the panel simulations (Fig. 3b, bottom row), both their mutational spectra and signature composition (bars below the spectra) are very similar. Moreover, Sig3 is dominant in Sig3⁺ cases and completely absent in Sig3⁻ cases; high-confidence cases have even greater presence (61 versus 37%) of Sig3. Although we do not have the gold-standard set of Sig3 MSK-IMPACT cases, the labels for the simulated panel data were derived from the WGS data. The similarity we observe indicates that our predictive model and the estimated sensitivity and specificity are applicable to the clinical panels. In a set of MSK-IMPACT cases with likely biallelic inactivation of *BRCA1/2*, we identified 75% to be Sig3⁺ (Supplementary Note), although the number of cases is not large enough (12) to be conclusive.

Furthermore, we investigated to what extent Sig3⁺ tumors exhibit copy number imbalance, which is a typical feature of HR-deficient tumors. Although the copy number profiles inferred from the panel data are much lower in resolution (Supplementary Note), our calculations show that high-confidence Sig3⁺ tumors have more imbalanced genomes than others (Fig. 3c, $P=10^{-5}$; one-sided Kolmogorov–Smirnov test). The genomic instability observed in the predicted Sig3⁺ tumors supports the validity of our approach.

Sig3⁺ cases in other tumor types. Although HR deficiency has been most closely associated with breast cancers, it can also manifest itself in cancers of other tissues, often through mechanisms that have not been clarified yet. For example, one possible mechanism for HR deficiency, which has been described recently, is the *EWS–FLI1* fusion in Ewing sarcomas[26]. This fusion leads to accumulation of R-loops that prevent the distribution of *BRCA1* to double-strand breaks, resulting in HR deficiency. Thus, in addition to those tumor types known to be associated with HR deficiency (for example, ovarian, uterine and pancreatic cancers), many other tissues may exhibit Sig3.

There are challenges in applying SigMA to the panel data from other tumor types. First, some tumor types have very low mutational burden, and panels do not capture a sufficient number of mutations for inference. For instance, none of the simulated panels for Ewing sarcomas and medulloblastomas had five or more mutations. For such tumor types, a larger panel or exome sequencing would be required to detect mutational signatures. Second, in some tumor types, other signatures that accompany Sig3 may generate most of the mutations. For example, in prostate tumors, Clock signatures are very active, making the detection of Sig3 more difficult. Finally, SigMA relies on the reference set of clusters (Fig. 1a) for classification (Supplementary Note). Until the number of publicly available WGS samples is sufficiently large, detection of Sig3 remains less

sensitive. Sensitivity ranges from 60% in osteosarcoma to 94% in pancreatic adenocarcinomas at an FPR of 10%. The FDR rates at a fixed FPR depend on the prevalence of HR deficiency in that tumor type (Supplementary Table 2).

Nonetheless, SigMA detects Sig3 from panels for multiple tumor types with a frequency ranging from 38% in ovarian cancers to 5% in esophageal carcinomas (Table 1). These values are obtained using the stringent settings of SigMA, with a 1% FPR in breast cancer and ranging between 1% and 5% in other tumor types, to provide a conservative lower bound on the use cases. For the tumor types associated with HR deficiency in the literature, that is for ovarian cancer, endometrial cancer, prostate adenocarcinoma and pancreatic cancer[27–31] SigMA identifies 38.0, 18.7, 12.5 and 7.1% of cases to be Sig3$^+$, respectively. For the other tumor types, such as Ewing sarcoma, osteosarcoma, medulloblastoma, bladder, esophageal and stomach adenocarcinoma, our results suggest that 5.0–27.3% are positive for Sig3.

Response to PARP inhibitors in Sig3$^+$ cell lines. To test the hypothesis that the presence of Sig3 indicates susceptibility to PARP inhibition, we examined the response of diverse cancer cell lines to four PARP inhibitors (olaparib, rucaparib, talazoparib and veliparib), using the Genomics of Drug Sensitivity in Cancer (GDSC) database[32]. We applied SigMA to mutation calls from a 1,651-gene capture panel from the Cancer Cell Line Encyclopedia (CCLE) project[33] to identify those with Sig3. In total, 383 cell lines from 14 tumor types were included in the study (Methods).

For each tumor type and PARP inhibitor, we first studied whether there was a difference in the sensitivity of Sig3$^+$ and Sig3$^-$ cell lines. Figure 4a displays the negative logarithm of the half-maximal inhibitory concentration ($pIC_{50} = -\log(IC_{50})$) for olaparib in 14 tumor types. We use pIC_{50} instead of IC_{50} to be less sensitive to small changes in the concentration, especially at low IC_{50} values. Of the 41 breast cancer cell lines, SigMA predicts 18 to be Sig3$^+$. The pIC_{50} values for olaparib (Fig. 4a) are significantly higher for the 18 Sig3$^+$ cell lines than for the 23 Sig3$^-$ cell lines (a positive shift of 1.5 in median pIC_{50}; $P = 0.0002$, t-test), indicating that Sig3 is associated with susceptibility to PARP inhibitors. For rucaparib, talazoparib and veliparib, the Sig3$^+$ breast cell lines also have higher pIC_{50} values, although the difference in pIC_{50} for Sig3$^+$ and Sig3$^-$ cell lines is smaller for rucaparib and veliparib (Fig. 4b). Across all the PARP inhibitors and tumor types examined, the median pIC_{50} in Sig3$^+$ cell lines was higher in the majority of cases (45 out of 56; Fig. 4a and Supplementary Fig. 15). For at least one of the PARP inhibitors, the pIC_{50} values are significantly higher ($P < 0.05$ at least) in Sig3$^+$ samples for Ewing's sarcoma, lymphoma, glioma, lung, ovarian, pancreatic and stomach cancer cell lines. For some other tumor types, such as osteosarcoma and prostate, the number of cell lines is too few for adequate power in tumor type-specific comparisons.

When data from all tumor types are combined (with appropriate normalization to account for different ranges of IC_{50} values; Supplementary Note), the normalized pIC_{50} values for olaparib are significantly higher for the Sig3$^+$ group compared to the Sig3$^-$ group (Fig. 4c, $P < 0.0001$). This also holds for rucaparib ($P < 0.0001$), talazoparib ($P < 0.0001$) and veliparib ($P = 0.0004$) (Fig. 4c). Removing the cell lines with *BRCA1/2* mutations (two cell lines with biallelic inactivation, three cell lines with a single-nucleotide polymorphism (SNP) or copy loss on a single allele) from this analysis did not change our conclusion. Restricting the tumor types to those that are reported in Table 1 also yields similar differences in the Sig3$^+$ and Sig3$^-$ groups (Supplementary Fig. 14e,h).

To ensure that the observed effect is specific to PARP inhibitors, we also examined other drugs as controls. In Fig. 4d, we compare the change in the IC_{50} for the Sig3$^+$ and Sig3$^-$ groups for all small-molecular drugs with available response data. Only monoclonal antibodies (one drug, cetuximab) and macromolecules (one drug, rTRAIL) are excluded in this comparison because their mechanism of action involves non-cell-autonomous components (for example, immune cell-mediated killing). The drugs are ranked according to the difference in the mean pIC_{50} values of the Sig3$^+$ and Sig3$^-$ groups, denoted as ΔpIC_{50}. Drugs are further grouped according their mechanism of action. PARP inhibitors are among the categories with the best responses in the Sig3$^+$group, together with DNA double-strand break-inducing agents, DNA cross-linkers and anthracyclines. Because the Sig3$^-$ group is largely composed of hormone-receptor-positive or human epidermal growth factor receptor 2 (HER2)-positive breast cancers, estrogen receptor modulators and HER2 inhibitors fall on the opposite side of the distribution. These results provide experimental evidence for the validity of our approach in identifying Sig3$^+$ cases and their sensitivity to PARP inhibitors, not only in breast and ovarian tumors but in many other tumor types, irrespective of the mutational status for *BRCA1/2*.

Platinum response in ovarian cancer. Finally, we study the effect of platinum treatments on patients with Sig3$^+$ ovarian cancer to demonstrate the efficacy of SigMA in identifying patients with ovarian cancer with HR deficiency on the basis of panels, since ovarian cancer patients harboring *BRCA1/2* mutations have been shown to have better survival[34]. In this analysis, we compare the overall survival among patients with *BRCA1/2* mutations, patients with high-confidence Sig3$^+$ without *BRCA1/2* mutations and Sig3$^-$ patients identified from the panels simulated using The Cancer Genome Atlas (TCGA) exome dataset. As shown in the Kaplan–Meier analysis (Fig. 5a), the overall survival for Sig3$^+$ patients ($n = 76$) without *BRCA* mutation is comparable to that of patients with a *BRCA1/2* mutation ($n = 19$), with a P value of 0.74 (log-rank test). Furthermore, the risk of death is 46% lower for wild-type Sig3$^+$ *BRCA1/2* patients than the Sig3$^-$ cases (Fig. 5b; hazard ratio for death, 0.54; 95% confidence interval (CI) 0.35–0.84; $P = 0.006$; Cox regression). The identification of Sig3 from the exome dataset of the same cohort leads to similar results: in wild-type *BRCA1/2* tumors, the hazard ratio for Sig3$^+$ versus Sig3$^-$ patients is 0.53 (95% CI 0.37–0.74; $P < 0.001$; Supplementary Fig. 16). From these results, we expect a similar favorable outcome in patients with Sig3 and those with *BRCA1/2* mutations.

Discussion

Although whole-exome sequencing and WGS are now commonplace for exploratory analysis, panel-based sequencing for profiling actionable mutations is predominant in routine clinical settings. In this study, we presented a tool designed to carry out mutational signature detection from panel sequencing data. With its likelihood-based approach, SigMA works well even when the mutation count is extremely low. Our simulated panel-based prediction of Sig3$^+$ cases faithfully recapitulates the WGS-based results, and our drug response data provide experimental support. As thousands of cancer cases are being profiled by panels at many hospitals[19] and more mutational signatures are characterized, our approach should be fruitful in identifying the mechanisms underlying the mutations and whether they may be amenable to existing therapies.

For breast cancer, PARP inhibitors have been given only to *BRCA1/2*-mutant cases. Our results indicate that it may be expanded to a larger group of patients, depending on the exposure to Sig3. Given that there were about 270,000 newly diagnosed breast cancer cases in 2018 (ref. [35]), around 13,500–27,000 (5–10%) cases may be attributed to inherited mutations of *BRCA1/2* (ref. [36]) Our analysis based on simulated data suggest that around twice that number of cases (27,000–54,000) may have the genomic footprint of HR deficiency (Sig3) without inherited mutations. PARP inhibitors might be a promising option for these patients. In ovarian cancer, PARP inhibitors have been used as a maintenance therapy after platinum-based chemotherapy, regardless of the *BRCA1/2* mutation status[29].

The general efficacy of PARP inhibitors in ovarian cancer regardless of germline mutation status is in accordance with the widespread deficiency in the HR pathway in ovarian cancer, as reflected in the prevalence of Sig3 (Fig. 1b). In addition, other reports have suggested that ovarian cancers with the evidence of HR deficiency may exhibit a more favorable response to PARP inhibitors, compared to those without the evidence of HR deficiency[37,38]. The genomic evidence of HR deficiency, including the presence of Sig3, could be a predictive biomarker for PARP inhibitor response that should be used in addition to *BRCA1/2* germline mutations. It would be worthwhile to investigate whether other cancer types with Sig3 could benefit from PARP inhibitor treatments.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0390-2.

Received: 11 July 2018; Accepted: 13 February 2019;
Published online: 15 April 2019

References

1. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
2. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
3. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).
4. Alexandrov, L. et al. The repertoire of mutational signatures in human cancer. Preprint at https://doi.org/10.1101/322859 (2018).
5. Burns, M. B. et al. APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **494**, 366–370 (2013).
6. Drost, J. et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science* **358**, 234–238 (2017).
7. Fedeles, B. I., Chawanthayatham, S., Croy, R. G., Wogan, G. N. & Essigmann, J. M. Early detection of the aflatoxin B_1 mutational fingerprint: a diagnostic tool for liver cancer. *Mol. Cell. Oncol.* **4**, e1329693 (2017).
8. Haradhvala, N. J. et al. Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat. Commun.* **9**, 1746 (2018).
9. Meier, B. et al. Mutational signatures of DNA mismatch repair deficiency in *C. elegans* and human cancers. *Genome Res.* **28**, 666–675 (2018).
10. Nik-Zainal, S. et al. The genome as a record of environmental exposure. *Mutagenesis* **30**, 763–770 (2015).
11. Ohno, M. et al. 8-oxoguanine causes spontaneous de novo germline mutations in mice. *Sci. Rep.* **4**, 4689 (2014).
12. Zou, X. et al. Validating the concept of mutational signatures with isogenic cell models. *Nat. Commun.* **9**, 1744 (2018).
13. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
14. Davies, H. et al. HRDetect is a predictor of *BRCA1* and *BRCA2* deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).
15. Polak, P. et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* **49**, 1476–1486 (2017).
16. Zámborszky, J. et al. Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. *Oncogene* **36**, 746–755 (2017).
17. Northcott, P. A. et al. The whole-genome landscape of medulloblastoma subtypes. *Nature* **547**, 311–317 (2017).
18. Sachs, N. et al. A living biobank of breast cancer organoids captures disease heterogeneity. *Cell* **172**, 373–386.e10 (2018).
19. Zehir, A. et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **23**, 703–713 (2017).
20. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).
21. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**, 3673–3675 (2015).
22. 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 Biol.* **17**, 31 (2016).
23. Kazanov, M. D. et al. APOBEC-induced cancer mutations are uniquely enriched in early-replicating, gene-dense, and active chromatin regions. *Cell Rep.* **13**, 1103–1109 (2015).
24. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
25. Frampton, G. M. et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* **31**, 1023–1031 (2013).
26. Gorthi, A. et al. EWS-FLI1 increases transcription to cause R-loops and block BRCA1 repair in Ewing sarcoma. *Nature* **555**, 387–391 (2018).
27. 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).
28. Fraser, M. et al. Genomic hallmarks of localized, non-indolent prostate cancer. *Nature* **541**, 359–364 (2017).
29. Ledermann, J. et al. Olaparib maintenance therapy in platinum-sensitive relapsed ovarian cancer. *N. Engl. J. Med.* **366**, 1382–1392 (2012).
30. Waddell, N. et al. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495–501 (2015).
31. Wu, Y.-M. et al. Inactivation of CDK12 delineates a distinct immunogenic class of advanced prostate cancer. *Cell* **173**, 1770–1782.e14 (2018).
32. Yang, W. et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–D961 (2013).
33. Basu, A. et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* **154**, 1151–1161 (2013).
34. Alsop, K. et al. *BRCA* mutation frequency and patterns of treatment response in *BRCA* mutation-positive women with ovarian cancer: a report from the Australian Ovarian Cancer Study Group. *J. Clin. Oncol.* **30**, 2654–2663 (2012).
35. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *CA Cancer J. Clin.* **68**, 7–30 (2018).
36. Roy, R., Chun, J. & Powell, S. N. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat. Rev. Cancer* **12**, 68–78 (2012).
37. Mirza, M. R. et al. Niraparib maintenance therapy in platinum-sensitive, recurrent ovarian cancer. *N. Engl. J. Med.* **375**, 2154–2164 (2016).
38. Telli, M. L. et al. Homologous-recombination deficiency (HRD) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple-negative breast cancer. *Clin. Cancer Res.* **22**, 3764–3773 (2016).

Acknowledgements

This work was mainly supported by the Ludwig Center at Harvard. I.C.C. received funding from the European Union (Marie Curie Skłodowska-Curie grant no. 703543). We would like to thank S. Elledge, G. Wulf, J. Dry and Z. Lai for helpful discussions, A. Galor and J. Cook for careful reading of the manuscript and S. Ouellette for help with the website.

Author contributions

D.C.G. and J.J.K.L. conceived the project. P.J.P. supervised the project. D.C.G. developed the algorithm, with suggestions and assistance from G.E.M.M., J.J.K.L. and I.C.C. In particular, G.E.M.M. helped with the simulation studies, J.J.K.L. suggested the application of signature analysis to PARP inhibitors and I.C.C. suggested the analysis of cell line/drug response data. D.C.G. and P.J.P. wrote the manuscript with input from all other authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0390-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to P.J.P.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019

Methods

SNV and somatic copy number alteration (SCNA) calls for tumors. The SNV and SCNA calls for WGS datasets from the TCGA project cohorts were downloaded from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/); those from the International Cancer Genome Consortium (ICGC) project were downloaded from its DCC data portal (https://dcc.icgc.org/releases). The somatic mutation datasets derived from WGS of 80 additional breast tumor normal pairs[14] were downloaded from the Department of Medical Genetics (S. Nik-Zainal) at the University of Cambridge (http://medgen.medschl.cam.ac.uk/serena-nik-zainal/). Consensus SNV and SCNA calls for the MSK-IMPACT panel data[19] were downloaded from the cBioPortal (http://cbioportal.org/msk-impact). SCNA calls for the MSK-IMPACT data were produced using CNVkit[39].

SNP and SCNA calls and drug response data for cell lines. SNP- and SNP-array-derived SCNA calls for the cancer cell lines from the CCLE[33] were downloaded from the CCLE data portal (https://portals.broadinstitute.org/ccle/). In vitro drug sensitivity information of the relevant cancer cell lines to various compounds, including PARP inhibitors from the GDSC database, were downloaded from https://www.cancerrxgene.org/ (ref. [32]).

SigMA. To detect mutational signatures from the SNV calls of whole-genome, exome or targeted gene panel data, we developed SigMA. A detailed description of the algorithm and its performance is provided in the Results and Supplementary Note.

In brief, SigMA consists of five main steps (see also the Supplementary Note): (1) mutational signatures in WGS data are discovered using NMF; (2) tumor subtypes according to their signature composition are determined with clustering and used as a reference for the panels; (3) we simulate cancer gene panels and exomes from the WGS data. In our simulations, the labels (whether a tumor is true $Sig3^+$ or $Sig3^-$) are known according to the signature analysis in the WGS data; (4) the likelihood measure, cosine similarity and exposure of Sig3 with NNLS are calculated for the simulated panels, exomes and WGS data; (5) we train gradient boosting classifiers specific for each tumor type, and sequencing platform, using the features from step 4. The gradient boosting classifiers yield a final combined score. We determine the thresholds on the SigMA score, which corresponds to small FPRs, using the simulated data and the true labels from the WGS analysis. The thresholds depend on tumor type and on the platform.

Clustering tumors by signatures to define tumor subtypes. Tumors are clustered according to the fractions of signature exposures and the existence of Sig3, a feature that takes a binary value (0 for tumors without Sig3 and 1 for tumors with Sig3), using hierarchical clustering (Fig. 1a and Supplementary Fig 3a). Tumors that are microsatellite stable without *POLE* mutations and without the effects of exogenous mutagens (for example, tobacco smoke, ultraviolet radiation) are clustered within each tumor type, while MSI-high tumors from all the tumor types are clustered together. To choose the number of clusters, the within-cluster and between-cluster sum of squares are considered. Once the clusters are defined, the average mutational spectrum of a cluster is calculated, first by normalizing the spectrum of each tumor to 1 and then taking the average of the normalized spectra of all the tumors in that cluster.

Likelihood calculation. The probability of observing a set of mutations for a given underlying mutational signature is calculated using Bayes' theorem and multinomial probability distributions. Multinomial distribution is a generalization of a coin tossing example (see Supplementary Note). Briefly, the number of mutations is equivalent to the number of times the coin is tossed, and the coin has 96 faces instead of 2. Trying to infer the underlying mutational signature from observed mutations is similar to attempting to tell which coin among several was tossed based on the observed head and tail counts. A formal description can be found in the Supplementary Note.

Simulations for tuning and testing the multivariate model. To tune the multivariate model and to test its performance, it is necessary to have a set of panels for which we know the true status of Sig3. In another study, where the HR deficiency was identified from the WGS data, Davies et al.[14] used the tumors with biallelic inactivation of *BRCA1/2* as a true positive set of HR deficiency. However, since HR deficiency is more prevalent than *BRCA1/2* mutations, we use the WGS NMF results as a reference, and we simulate targeted sequencing panels from the WGS data. The simulations are done by downsampling the WGS data to the target regions of the panels. However, we found that the difference in depth of coverage between WGS (approximately 40×) and panel sequencing (approximately 1,000×) resulted in a smaller number of mutations in our simulated panel, compared to the original panel datasets. Therefore, we increased the number of mutations in the panel simulations by randomly sampling the WGS data. The amount of additional mutations we added in this way and how we determined the effects of differences in coverage are discussed in the Supplementary Note.

Drug response in cell line models. Mutation calls from a 1,651-gene capture panel and copy number calls from SNP arrays were available for each cell line from the CCLE project; the exome sequences of the same cell lines are available independently from the GDSC project. However, in this analysis we did not use the whole exomes from the GDSC project, due to the differences in the mutational spectra between the CCLE and the GDSC data (Supplementary Fig. 22a,b). The spectra of simulated MSK-IMPACT panels from the WGS data of tumors were more similar to the CCLE results. The different mutational spectra between the two studies (higher C>A and T>G in the GDSC mutation calls compared to those from the CCLE) are not explained by the different trinucleotide frequencies in the target regions of the platforms used in those studies. Among 1,534 cell lines in total, 874 had drug response data for either one of the four PARP inhibitors. In our main analysis, we used 383 out of 874 cell lines, excluding those that: (1) had few samples available, such as uterine and medulloblastoma cell lines, both of which had only two cases with PARP inhibitor response; (2) had no $Sig3^+$ cases; (3) belonged to tumor types for which we have not attempted a Sig3-based analysis, such as liver and kidney; or (4) belong to tumor types for which no WGS dataset was publicly available, such as neuroblastoma. We used a stringent filter to discriminate the cell lines with Sig3 mutations from those without Sig3 but with a large number of in vitro culture-associated mutations (Methods and Supplementary Note). We also removed cell lines with MSI signatures to minimize their confounding effect (see Vilar Sanchez et al.[40] and Supplementary Fig. 14c). More information on cell line analysis can be found in the Supplementary Note.

Survival study for patients with ovarian cancer, based on Sig3 status. The clinical data for patients with ovarian cancer from the TCGA consortium were obtained from the Broad Institute Genomic DNA Affinity Chromatography Firehose (https://gdac.broadinstitute.org/). Cases treated with cisplatin and carboplatin were selected. The exome data from the TCGA consortium were used to generate the simulated panels, which had a mutational burden in agreement with that in MSK-IMPACT data for ovarian cancer. The tumors with *BRCA* mutations are separated from the rest of the cohort. The high-confidence $Sig3^+$ cases were identified with SigMA, and the overall survival chances of these patients are shown in comparison to the rest. Survival analysis was carried out using the 'survminer' and 'survival' R packages. More information on the survival analysis can be found in the Supplementary Note.

Statistics. For the Kolmogorov–Smirnov and *t*-tests, the 'stats' R package was used. The log-rank and likelihood-ratio tests used in the survival analysis are implemented within the survminer R package.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Detailed information on how to access the ICGC, TCGA, CCLE and GDSC data for the cell lines can be found in the Methods. Information about the ICGC and TCGA can be found at https://icgc.org/ and http://cancergenome.nih.gov, respectively All other remaining data are available within the article and in the Supplementary Data, or available from the authors upon request.

Code availability

The code for SigMA is available on GitHub (https://github.com/parklab/SigMA) as an R package.

References

39. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput. Biol.* **12**, e1004873 (2016).
40. Vilar Sanchez, E. et al. Preclinical testing of the PARP inhibitor ABT-888 in microsatellite instable colorectal cancer. *J. Clin. Oncol.* **27**, 11028 (2009).

nature research

Corresponding author(s): Peter J. Park

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	This study does not include any new data collection, therefor no software was used.
Data analysis	The code will be made public in a github repository, https://github.com/parklab/SigMA.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data availability statement is at the end of the manuscript in Online Methods section.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Size was chosen based on the sample size in the publicly available datasets. The sample size for each cancer type was estimated by the International Cancer Genome Consortium (ICGC), CCLE and GDSC projects, and Memorial Sloan Kettering Cancer Center.
Data exclusions	We only considered sequencing data passing the Quality & Control criteria established by the ICGC and Pan-Cancer Analysis of Whole Genomes Project (PCAWG), and CCLE and GDSC Projects,Memorial Sloan Kettering Cancer Center. The quality controls on each of these public datasets are discussed in their corresponding publication. We state this in the main manuscript.
Replication	In order to allow reproducibility of our work, and to avoid over-fitting of our models, we use cross-validation in the training of machine learning methods to select the optimal parameters (Supplementary Notes Section 10). We split data into non-overlapping test and training datasets using stratified folding, when we calculate the false positive rate and sensitivity quoted in Figure 2. We also validated our model, which was optimized using simulations of panels from whole genome sequenced ICGC and PCAWG samples, on independent datasets such as published MSK-IMPACT data and cell line models.
Randomization	After ordering the data according to signal strength we randomly selected the training samples and test samples, which allows an unbiased selection of signal strength in test and training datasets.
Blinding	This study does not include any data collection and public data is used. There are no instances of grouping based on the choice of the investigators. We determine Signature 3-positive and -negative samples based on the signature analysis we carry out in the WGS data. We determine test and training sample with stratified folding procedure. For other analysis the true labels of patients were not known when the classification was done, i.e. for MSK-IMPACT panels the truth labels are not known.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	The co-variates in the overall survival study for ovarian cancer patients were included as effect parameters in the Cox regression. The covariates are listed in Figure 5. The datasets obtained from TCGA and ICGC, which are used as a reference in training of our model, are assumed to represent the general population of patients with different types of cancer. This might not be strictly true for tumor types with small number of tumors such as bladder cancer, and the results should be used with caution.
Recruitment	The whole genome sequenced samples were obtained from participants were recruited in each individual ICGC study, in TCGA., and for panel data by MSK center. All these studies complied with the required ethical guidelines.