RESEARCH ARTICLE

WILEY

Analysis of mutational signatures with yet another package for signature analysis

Daniel Hübschmann[1,2,3,4] | Lea Jopp-Saile[1,2,5] | Carolin Andresen[2,5] |
Stephen Krämer[5] | Zuguang Gu[1,6] | Christoph E. Heilig[3,7] |
Simon Kreutzfeldt[3,7] | Veronica Teleanu[3,7] | Stefan Fröhling[3,7] | Roland Eils[8,9] |
Matthias Schlesner[5,10]

[1]Computational Oncology, Molecular Diagnostics Program, National Center for Tumor Diseases (NCT) and German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, Heidelberg, 69120, Germany

[2]Pattern Recognition and Digital Medicine, Heidelberg Institute of Stem Cell Technology and Experimental Medicine (HI-STEM), Im Neuenheimer Feld 280, Heidelberg, 69120, Germany

[3]German Cancer Consortium (DKTK), Im Neuenheimer Feld 280, Heidelberg, 69120, Germany

[4]Department of Pediatric Immunology, Hematology and Oncology, University Hospital Heidelberg, Heidelberg, Germany

[5]Bioinformatics and Omics Data Analytics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, Heidelberg, 69120, Germany

[6]DKFZ-HIPO (Heidelberg Center for Personalized Oncology), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, Heidelberg, 69120, Germany

[7]Department of Translational Medical Oncology, NCT Heidelberg and DKFZ, Heidelberg, Germany

[8]Center for Digital Health, Berlin Institute of Health and Charité – Universitätsmedizin Berlin, Berlin, Germany

[9]Health Data Science Unit, University Hospital Heidelberg, Heidelberg, Germany

[10]Chair of Biomedical Informatics, Data Mining and Data Analytics, Faculty of Applied Computer Science and Medical Faculty, University of Augsburg, Augsburg, Germany

Correspondence
Daniel Hübschmann MD, PhD, CASM, Computational Oncology, Molecular Diagnostics Program, National Center for Tumor Diseases (NCT) and German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany.
Email: d.huebschmann@dkfz-heidelberg.de

Matthias Schlesner PhD, Chair of Biomedical Informatics, Data Mining and Data Analytics, Faculty of Applied Computer Science and Medical Faculty, University of Augsburg, Augsburg, Germany.
Email: matthias.schlesner@informatik.uni-augsburg.de

Abstract

Different mutational processes leave characteristic patterns of somatic mutations in the genome that can be identified as mutational signatures. Determining the contributions of mutational signatures to cancer genomes allows not only to reconstruct the etiology of somatic mutations, but can also be used for improved tumor classification and support therapeutic decisions. We here present the R package yet another package for signature analysis (YAPSA) to deconvolute the contributions of mutational signatures to tumor genomes. YAPSA provides in-built collections from the COSMIC and PCAWG SNV signature sets as well as the PCAWG Indel signatures and employs signature-specific cutoffs to increase sensitivity and specificity. Furthermore, YAPSA allows to determine 95% confidence intervals for signature exposures, to perform constrained stratified signature analyses to obtain enrichment and depletion patterns of the identified signatures and, when applied to whole exome sequencing data, to correct for the triplet content of individual target capture kits. With this

functionality, YAPSA has proved to be a valuable tool for analysis of mutational signatures in molecular tumor boards in a precision oncology context. YAPSA is available at R/Bioconductor (http://bioconductor.org/packages/3.12/bioc/html/YAPSA.html).

KEYWORDS
biomarker, BRCAness, cancer genomics, mutational signatures, precision oncology, YAPSA

1 | INTRODUCTION

Mutational signatures are patterns of mutations arising from specific mutational processes. Point mutations (single nucleotide variants, SNVs) in their trinucleotide motif context[1] have been studied intensively in this regard. Mutational signature analysis can not only inform about the etiology of mutations,[1] but also give insights into tumor evolution,[2] improve diagnosis and patient stratification[3] and support therapy decision-making.[4]

The de novo identification of mutational signatures depends on the availability of very large series of cancer whole genome sequencing (WGS) or whole exome sequencing (WES) data. In an initial unsupervised de novo analysis of 507 WGS and 6535 WES tumor samples from 30 different tumor entities, Alexandrov et al. identified 27 mutational signatures.[1] An extended analysis of 10 952 WES samples and 1048 WGS samples across 40 different tumor entities revealed 30 validated mutational signatures https://cancer.sanger.ac.uk/cosmic/signatures_v2. Recently, the pan-cancer analysis of whole genomes (PCAWG) consortium has published two even larger collections of SNV mutational signatures extracted from 4645 WGS and 19 184 WES tumor samples from 92 cancer entities computed by two different non-negative matrix factorization (NMF) algorithms.[5] A consensus of 67 signatures, 47 of which are termed "likely to be real", between these two collections was proposed (https://cancer.sanger.ac.uk/cosmic/signatures_v3). Furthermore, the PCAWG consortium has provided a classification system to group small insertions and deletions (Indels) into 83 features and has extracted 17 Indel mutational signatures (https://cancer.sanger.ac.uk/cosmic/signatures/ID) from the subcohort of 2780 WGS tumor samples across 37 entities.[5]

For roughly half of the identified mutational signatures, underlying mutational processes have been assigned.[1,5] These may by grouped into (a) those linked to aging (clock-like signatures, for example, spontaneous deamination), (b) those related to the action of exogenic carcinogens (eg, benzoapyrene from tobacco smoke or UV light), (c) those related to defects in DNA repair pathways (eg, homologous recombination repair (HRR) or mismatch repair (MMR)) or (d) those related to over-activation of physiologically mutagenic enzymes (APOBEC or AID).[1,5]

If a consensus set of mutational signatures is already known, de novo extraction is not necessary. Instead, the contributions of (known) signatures can be computed in a supervised analysis or fitting of mutational signatures.[6-11] This allows to determine mutational signature contributions in small sample sets and even individual samples,

thereby enabling the use of mutational signatures to improve cancer diagnosis and as biomarkers for therapy sensitivity and resistance prediction (reviewed in Van Hoeck et al[3]). However, fitting of mutational signatures can lead to false negatives (mutational processes in the studied cohort that are missing in the used signature set) and false positives (erroneous detection of mutational processes, which have not been active in the respective samples, possibly due to overfitting).[12] While the occurrence of false negatives is reduced with the availability of increasingly comprehensive sets of mutational signatures, false positives can only be avoided by filtering or correction strategies in the deconvolution process.

The R/Bioconductor package yet another package for signature analysis (YAPSA) presented herein performs fitting of mutational signatures. It can determine contributions for both SNV and Indel mutational signatures and for both the COSMIC and PCAWG signature collections. YAPSA employs a filtering strategy with signature-specific cutoffs to reduce false positive signature calls and thereby increases specificity and precision. Furthermore, YAPSA provides 95% confidence intervals (CIs) for the determined signature contributions and offers the possibility to run constrained stratified analysis to find enrichment and depletion patterns of the identified signatures in subsets of the called somatic mutations.

Precision oncology subjects samples from individual patients to broad genomic profiling to identify clinically actionable lesions.[13-20] YAPSA may be particularly well suited for application in precision oncology due to its high specificity achieved through signature-specific cutoffs and to the possibility to assign confidence levels to signature calling. Since 2015, YAPSA has been used in a registry trial for younger adults with advanced cancer across histologies and patients with rare tumors across all age groups.[21,22]

2 | MATERIALS AND METHODS

2.1 | General setting and nomenclature

In this work, according to a widely used convention, we call the matrix of the counted occurrences of every feature (in the case of SNV mutational signatures the features are SNVs in their triplet context) in every sample the mutational catalog V. It is the aim of the mutational signature analysis to decompose this mutational catalog V into two smaller matrices W and H such that

$$V \approx W \cdot H \qquad (1)$$

The columns of the matrix *W* are the mutational signatures and the columns of the matrix *H* are the contributions of every signature to every sample. *W* is called the signature matrix and *H* is called the exposure matrix.

In an unsupervised extraction or de novo analysis of mutational signatures, both *W* and *H* are unknown. As opposed to that, when fitting mutational signatures, *W* is known and only *H* is unknown and has to be determined. We denote the *j*-th column of *V* by $V(\cdot;j)$, corresponding to the mutational catalog of sample *j*. Analogously we denote the *j*-th columns of *H* by $H(\cdot;j)$, which is the exposure vector of sample *j*. Then the task of finding *H* can be written as an optimization problem:

$$\min_{H_{(\cdot;j)} \in \mathbb{R}^l} \left\| W \cdot H_{(\cdot;j)} - V_{(\cdot;j)} \right\| \ \forall j \in \{1,...,m\}, \qquad (2)$$

Under the constraint of non-negativity:

$$H_{(ij)} \geq 0 \ \forall i \in \{1,...,l\} \forall j \in \{1,...,m\}. \qquad (3)$$

The number of samples is *m*, *i* is the index over signatures, *l* is the number of signatures, and *j* is the index over samples.

In YAPSA, the task of finding *H* is accomplished by a family of functions called `LCD()` (linear combination decomposition). These functions run a three-step procedure:

1. a non-negative least squares (NNLS) using the function `nnls()` from the R package nnls[23] is performed
2. for all signatures, the obtained exposures, that is, the contributions of the signatures to the overall mutational catalog, are compared to the respective signature-specific cutoffs, and only those signatures for which the exposures are higher than the respective optimal signature-specific cutoffs are kept
3. The NNLS procedure is re-run with only the remaining signatures.
4. After having computed the exposures, the user can in addition compute 95% CIs by using the functions `variateExp()` (for SNV mutational signatures) or `confidence_indel_only_calculation()` (for Indel mutational signatures).

For a given cohort, fitting of mutational signatures can be performed at cohort-wide or per-sample level. In a cohort-wide analysis, the NNLS is applied to the whole mutational catalog (i.e., the whole matrix *V* with as many columns as there are samples in the cohort) simultaneously, whereas in a per-sample analysis, the mutational catalog is split into the different columns (the individual patients) and the NNLS is applied to the different columns separately. In YAPSA, the function `LCD_complex_cutoff_combined()` performs both analyses and the user can access the results of these two complementary analyses separately.

2.2 | Definition of features and signatures

For an analysis of mutational signatures, the mutations are grouped into categories or features.

2.2.1 | SNV signatures

For SNV mutational signatures, these features are determined by the nucleotide exchange and the surrounding triplet motif context resulting in 96 features.[1] Using lists of SNV variant calls, in YAPSA the SNV mutational catalog is assembled by the functions `create_mutation_catalogue_from_df()` or `create_mutation_catalogue_from_VR()`, which use the functions `mutationContext()` and `motifMatrix()` from the bioconductor package SomaticSignatures.[24]

2.2.2 | Indel signatures

According to the classification provided by the PCAWG consortium, Indels are categorized into 83 features in 16 groups.[5]

- Groups 1 and 2: Deletions of 1 bp C/(G) or T/(A): features are classified by the repetitive context in which the deletion occurs: 1, 2, 3, 4, 5, or larger than or equal to 6 times the same nucleotide.
- Groups 3 and 4: Insertions of 1 bp C/(G) or T/(A): features are classified by the repetitive context in which the insertion occurs: 0, 1, 2, 3, 4, or larger than or equal to 5 times the same nucleotide. Here zero represents an inserted motif in a sequence context where it was not present before.
- Groups 5 to 8: Deletions of 2 bps, 3 bps, and 4 bps or more than or equal to 5 bps: features are classified by the repetitive context in which the deletion occurs: 1, 2, 3, 4, 5 or larger than or equal to 6 times the same deleted motif.
- Groups 9 to 12: Insertions of 2 bps, 3 bps, and 4 bps or more than or equal to 5 bps: features are classified by the repetitive context in which the insertion occurs: 0, 1, 2, 3, 4 or larger than or equal to 5 times the same inserted motif. Here zero represents an inserted motif in a sequence context where it was not present before.
- Groups 13 to 16: Deletions of 2 bps, 3 bps, 4 bps or more than or equal to 5 bps with microhomology at the breakpoints: features are defined by the loss of a motif in a partially receptive context of 1 bp, 2 bps, 3 bps, 4 bps or more than or equal to 5 bps either 5′ or 3′ of the deletion.

Using the above nomenclature, for Indel mutational signatures, the mutational catalog *V* has 83 rows. In YAPSA, the categorization and counting of Indel mutations and the generation of an Indel mutational catalog is performed by the function `create_indel_mutation_catalogue_from_df()`. The evaluation of microhomology for attribution of a given Indel to Groups 13 to 16 is performed by the function `matchPattern()` from the R package Biostrings.[25]

2.3 | Signature-specific cutoffs

YAPSA provides optimal signature-specific cutoffs in order to filter the signatures (step 2 in Section 3.1) identified by an initial NNLS computation (step 1 in Section 3.1).

These optimal signature-specific cutoffs were trained by an optimizing procedure using a modified Receiver Operating Characteristic (ROC) analysis implemented in the R package ROCR.[26] The objective of the ROC was to maximize the overlap of the signatures identified by the YAPSA analysis in its filtering step with the signatures identified in the original de novo NMF analysis of the respective data sets (Alexandrov et al., 2013[1] for COSMIC SNV signatures and Alexandrov et al., 2020[5] for PCAWG signatures).

For COSMIC SNV signature analysis, data was downloaded from ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl/somatic_mutation_data/ on November 4, 2015. From every entity-specific subdirectory of this URL, files with the suffix "_clean_somatic_mutations_for_signature_analysis.txt", which contain the somatic SNV variant calls, were downloaded. The reference information which signature is present in which entity was extracted from the matrix displayed in the figure available at https://cancer.sanger.ac.uk/signatures_v2/matrix.png.

For PCAWG SNV and Indel signature analysis, data was downloaded from https://www.synapse.org/#!Synapse:syn11726601/files/ on October 9, 2019. Somatic variant calls were downloaded from the subfolder structure "Input_Data_PCAWG7_23K_Spectra_DB/ Mutation Catalogs – Spectra of Individual Tumors". The reference information per sample was downloaded from the subfolder structure "Signatures_in_Samples/SP_Signatures_in_Samples/".

2.4 | Confidence intervals

YAPSA uses the concept of profile likelihoods, which has been developed for modeling ordinary differential equations (ODEs),[27] to compute CIs for the exposures to mutational signatures. This is carried out by a multi-step procedure:

1. In YAPSA, when performing an analysis of mutational signatures with functions from the LCD() family, the signatures present in the data to be analyzed are determined in a filtering step (cf. Sections "4. Results" and "3.1 General Setting and Nomenclature"). For such a solution of equation (Equation (2) with exposures to a determined subset of l signatures { $W_{(\cdot;ind_1)}$, ..., $W_{(\cdot;ind_l)}$ }, the distribution of the residuals is determined, and then log-likelihoods are computed by the function variateExp(). In detail, this is performed as follows: for a given sample v, let $R_{(\cdot;v)}$ denote the v-th column of the matrix of residuals $R = W \cdot H - V$. Let $pdf()$ denote the probability distribution function of this vector of residuals. In YAPSA the default for $pdf()$ is set to be a normal distribution with identical mean and SD as $R_{(\cdot;v)}$. Alternatively, the user can also specify another pdf(). The YAPSA function computeLogLik() then calculates the log-likelihood:

$$LogLik_v = \sum_{j=1}^{l} log_2\left(pdf\left(R_{jv}\right)\right). \qquad (4)$$

2. We call this the *initial model* of the data. When computing the CI for the exposure H_{uv} to a given signature $W_{(\cdot;u)}$ in a sample v (using

the nomenclature defined above to denote the u-th column of the matrix W), this exposure is perturbed, that is, ., $H_{uv} \longrightarrow H_{uv} + \delta_{uv}^0$. Here, δ_{uv}^0 is the starting value for an iterative procedure (cf. below) and by default is set to $\delta_{uv}^0 = \pm 0.4$.

3. Then the exposures to the remaining l-1 signatures are computed again by NNLS:

$$\min_{H_{(1,...,u-1,u+1,...,l;v)} \in \mathbb{R}^{l-1}} \left(\left\| W_{(\cdot;1,...,u-1,u+1,...l)} \cdot H_{(1,...,u-1,u+1,...,l;v)} \right. \right.$$
$$\left. \left. - V_{(1,...,u-1,u+1,...,l;v)} \right\| + C \right). \qquad (5)$$

3. Where $C = \left| W_{(\cdot;u)} \cdot \left(H_{uj} + \delta_{uv}^0 \right) - V_{uv} \right|$. That leads to an alternative model of the data. The degrees of freedom of the *alternative model* are l-1, that is, one less than those of the *initial model*. Analogously to the *initial model* of the data, the log-likelihood $LogLik_v^{\#}$ is computed from the distribution of the residuals of this *alternative model*. Finally, a likelihood ratio test is computed using the log-likelihoods of the *initial* and *alternative models* by first computing a test statistic

$$TestStat = 2 \cdot \left(LogLik_v - LogLik_v \right), \qquad (6)$$

2. Then, using the R function pchisq(), a P value for this perturbation δ_{uv}^0. The perturbations δ_{uv}^0 may not necessarily correspond to two-sided 95% CIs, but instead to intervals relating to the computed P values of the likelihood ratio test. Therefore the Gauss-Newton method (function newtonsys() in the R package pracma[28]) is used to approximate actual 95% CIs.

If the Gauss-Newton method does not converge, the starting value for the perturbation is adjusted by $\delta_{uv}^0 \rightarrow \delta_{uv}^0 \cdot 2$ and steps (2) to (5) are re-run. If δ_{uv}^0 reaches a value ≥10, this outer iteration is stopped and no confidence interval can be computed.

2.5 | Application to WES data

WGS and WES differ in the frequency of occurrence of different k-mers. According to the concept underlying SNV mutational signatures,[1,5] the triplet (or 3-mer) context of an SNV is used for categorization of the mutations, leading to 96 different categories or features. The relative occurrence of these 96 different features differs between WGS and WES. More precisely, let n_X^{WGS} denote the occurrence of feature X in the whole genome and n_X^{WES} denote the occurrence of X in an exome target capture. We then define

$$q_X^{WGS,WES} = \frac{n_X^{WGS}}{n_X^{WES}}, \qquad (7)$$

to be the ratio of these two counts. These ratios are not identical for all features, that is, .,

$$\exists X, Y \in \mathbb{F} : q_X^{WGS,WES} \neq q_Y^{WGS,WES}, \qquad (8)$$

Where $\mathbb{F}$ denotes the feature space. It is thus crutial to compute $q_X^{WGS,WES}$ for all features $X \in \mathbb{F}$ and to correct for these differences. These corrections can be applied either to the signatures, converting them to "exome signatures", or the inverse corrections can be applied to the mutational catalogs. In YAPSA, we opt for the second alternative, as this keeps the function calls simple, analogous, and very similar for analyses of both WES and WGS data. Detailed information can be found in a vignette dedicated to this topic: https://www.bioconductor.org/packages/release/bioc/vignettes/YAPSA/inst/doc/vignette_exomes.html

2.6 | Stratified analysis of mutational signatures

YAPSA performs stratified analysis of mutational signatures with the constraints that (a) only those signatures, which have been detected in the entire set, may be used in the deconvolutions of the strata and (b) the sum of the exposures of a given sample over all strata must equal to the exposures of this sample in the unstratified analysis. A stratified analysis of mutational signatures starts by assigning the mutations to the different strata and building separate mutational catalogs V^k for these strata. The strata have to be provided by the user based on external criteria, for example, mutation density, replication timing, or chromatin states. The strata have to be exclusive, that is, one mutation cannot be in more than one stratum simultaneously. The constrained stratified analysis of mutational signatures is then performed by the function SMC() (Stratification of the Mutational Catalog). It solves the following optimization problem:

$$\min_{H_{(j)}^k \in \mathbb{R}^l} \left\| W \cdot H_{(j)}^k - V_{(j)}^k \right\| \ \forall j, k. \qquad (9)$$

under the constraint of non-negativity: $H_{(ij)}^k \geq 0 \ \forall i,j,k$. In addition, the additional constraint:

$$\sum_{k=1}^{s} H^k = H, \qquad (10)$$

Where H is defined as a solution to the unstratified optimization problem in equation (Equation (2)) and underlies the constraint of non-negativity, that is, H is a solution computed by LCD().

Naturally, the mutational catalogs of the strata have to sum up to the overall mutational catalog:

$$V = \sum_{k=1}^{s} V^k \qquad (11)$$

j is the index over samples, i is the index over signatures, l is the number of signatures, k is the index over strata, and s is the number of strata. Equation (Equation (11)) reflects the additivity of the stratified mutational catalogs V^k.

The SMC procedure can also be applied when the unstratified analysis was performed by another strategy, for example, an unsupervised de novo analysis or extraction of mutational signatures using NMF. Let $\tilde{H}$ denote the exposures of this preceding other analysis. SMC() then solves the task described by equation (Equation 9) with a slight difference in the additional constraint:

$$\sum_{k=1}^{s} H^k = \tilde{H}. \qquad (12)$$

Applying SMC() that way, the initial LCD() decomposition of the unstratified mutational catalog is omitted and its result replaced by the exposures of the preceding other analysis.

Further information about stratified analyses of mutational signatures can be found in the corresponding vignette: https://bioconductor.org/packages/3.12/bioc/vignettes/YAPSA/inst/doc/vignette_stratifiedAnalysis.html

2.7 | External data

For the data used to generate Figure 1, we refer to the Supplementary Information Section "3 External data". In this work, the functionality of the software package YAPSA is also demonstrated on an ovarian cancer data set.[29] Data was downloaded on the 4 June 2020 from https://dcc.icgc.org/api/v1/download?fn=/current/Projects/OV-AU/simple_somatic_mutation.open.OV-AU.tsv.gz. Only those samples with more than 25 SNVs and more than 20 Indels were kept for analysis. The whole analysis used to generate Figures 2–5 and Supplementary Figures 4-6 as well as all numeric values for this cohort are provided in the supplementary files *Code_for_figure_generation.Rmd* (R markdown) and *Code_for_figure_generation.html* (compiled report).

2.8 | Processing of samples from the precision oncology program MASTER

Two samples presented in this work underwent paired-end WGS (2 × 151 bp) on a HiSeq X instrument (Illumina, San Diego, California) in the framework of the MASTER (Molecularly Aided Stratification for Tumor Eradication Research) program of NCT Heidelberg/Dresden and the German Cancer Consortium is a registry trial for younger adults with advanced cancer across histologies and patients with rare tumors across age groups.[21,22] Library preparation was performed with the TruSeq Nano Library Preparation Kit (Illumina). Alignment with BWA mem, small variant calling, and calling of somatic copy number aberrations (sCNAs) with ACEseq[30] were performed as described earlier.[4,31,32] Somatic SNV (*case1_somaticSnvs.vcf*, *case2_somaticSnvs.vcf*) and somatic Indel (*case1_somaticIndels.vcf*, *case2_somaticIndels.vcf*) variant calls required for reproducibility of the analysis of mutational signatures as well as segment information obtained from sCNA calling
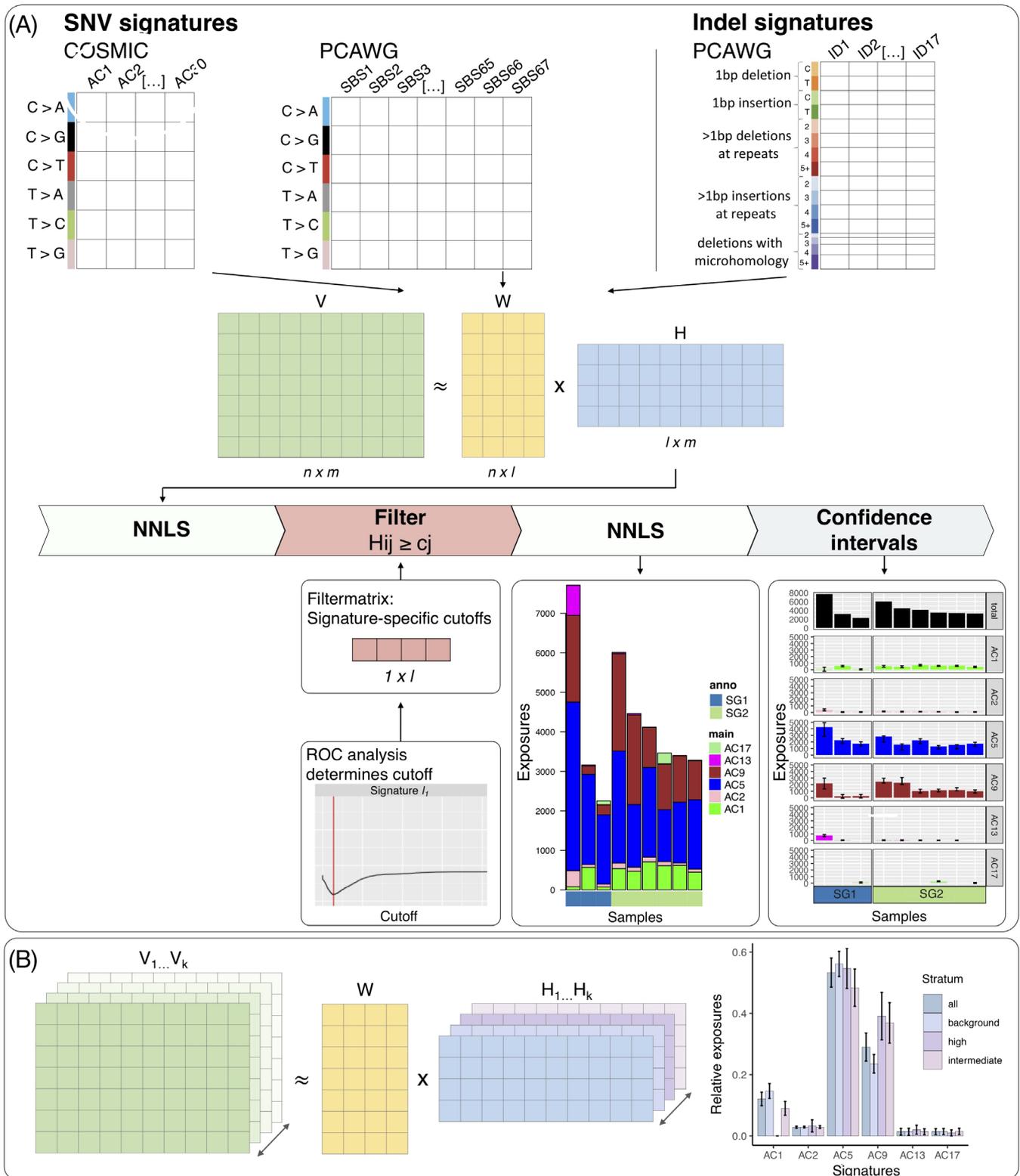
FIGURE 1 Functionalities of YAPSA. A, Contributions of mutational signatures (*H*) are computed for given mutational catalog (*V*) and signatures (*W*) by a multi-step procedure (central flow diagram). Results are visualized as (stacked) barplot; subgroup (SG) information can be used to pre-sort the samples. B, Stratification of the mutational catalog showing enrichment and depletion patterns of mutational signatures. The mutational catalog is decomposed into different strata ($V_1, \ldots V_k$) and with known signatures, *W* the task is to compute a collection of exposure matrices ($H_1, \ldots H_k$) under constraints. On the right, the y-axis represents normalized contributions of the different signatures in the respective strata. Error bars: SE of the mean (SEM). SG1, SG2: subgroup 1 or 2. SNV, single nucleotide variants; YAPSA, yet another package for signature analysis
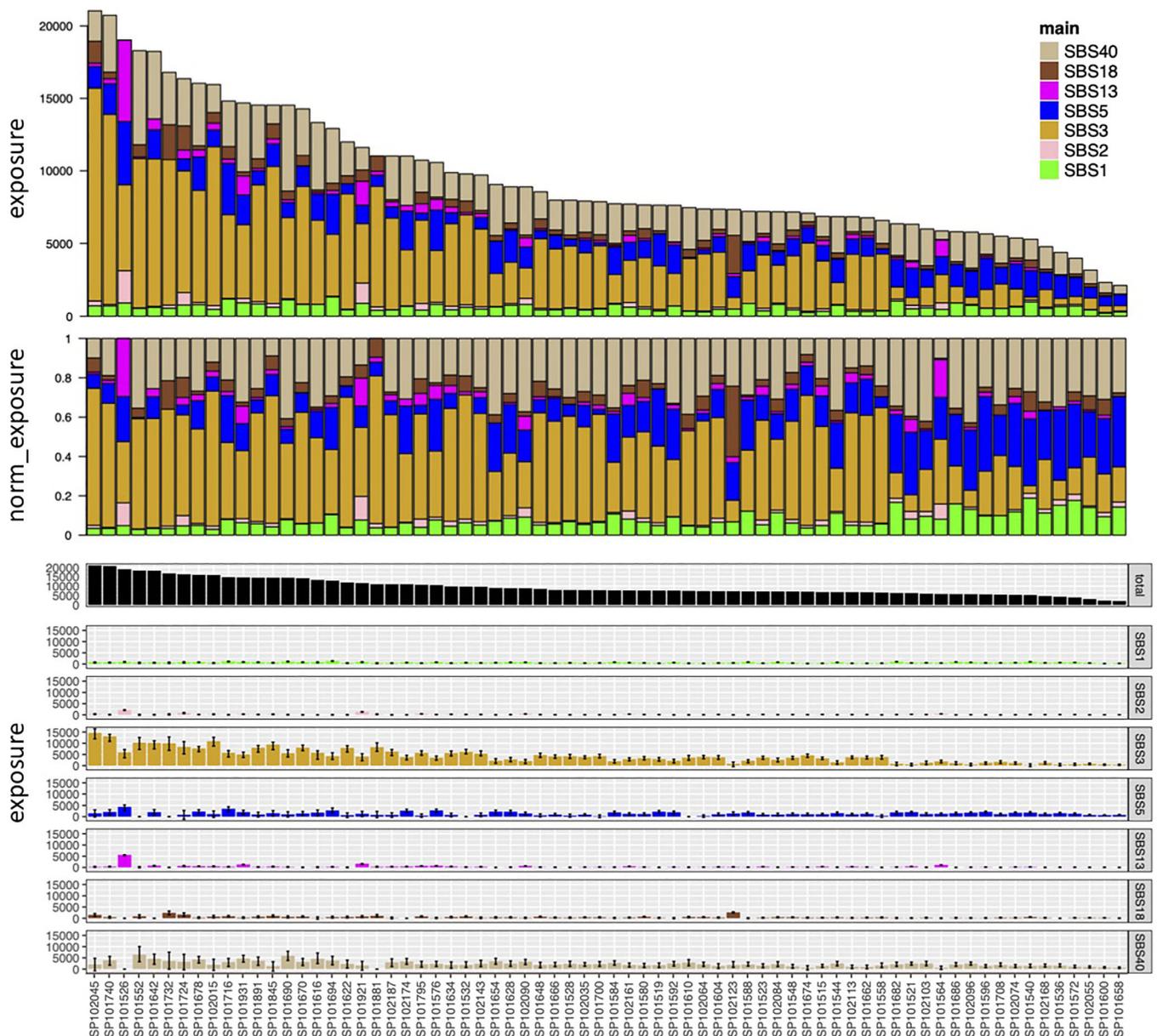
FIGURE 2 Cohort-wide analysis of an ovarian cancer cohort with PCAWG SNV signatures. Top panel: absolute exposures; the y-axis displays the number of SNVs explained by the respective mutational signatures. Middle panel: normalized exposures. Bottom panel: absolute exposures with 95% CIs; the y-axis displays the number of SNVs explained by the respective mutational signatures. Samples in all panels are ordered on the x-axis by decreasing SNV mutational load. SNV, single nucleotide variants

(*case1_segments.csv*, *case2_segments.csv*) required for the reproducibility of the computation of genomic instability (cf. below) are provided as supplementary files.

2.9 | Quantification of genomic instability

As described previously,[4] genomic instability was assessed using the Loss-Of-Heterozygosity-Homologous-Recombination-Deficiency (LOH-HRD) score[33-35] and the number of large-scale state transitions (LSTs).[36] Both measures rely on analysis and calling of sCNAs followed

by a reduction of oversegmentation due to technical noise by smoothing of the copy number profiles. Small segments, that are segments smaller than 3 Mbp, were processed as follows: (i) among the neighboring segments, the one more similar with respect to total and allele-specific copy number states was determined, and then (ii) the small segment was merged with the more similar neighbor. The LOH-HRD score corresponds to the number of subchromosomal segments with loss of heterozygosity larger than 15 Mbp.[33] Using the same smoothing step as the calculation of the LOH-HRD score, LSTs are defined to be switches between segments of different copy number states larger than 10 Mbp but smaller than entire chromosome arms.[36]
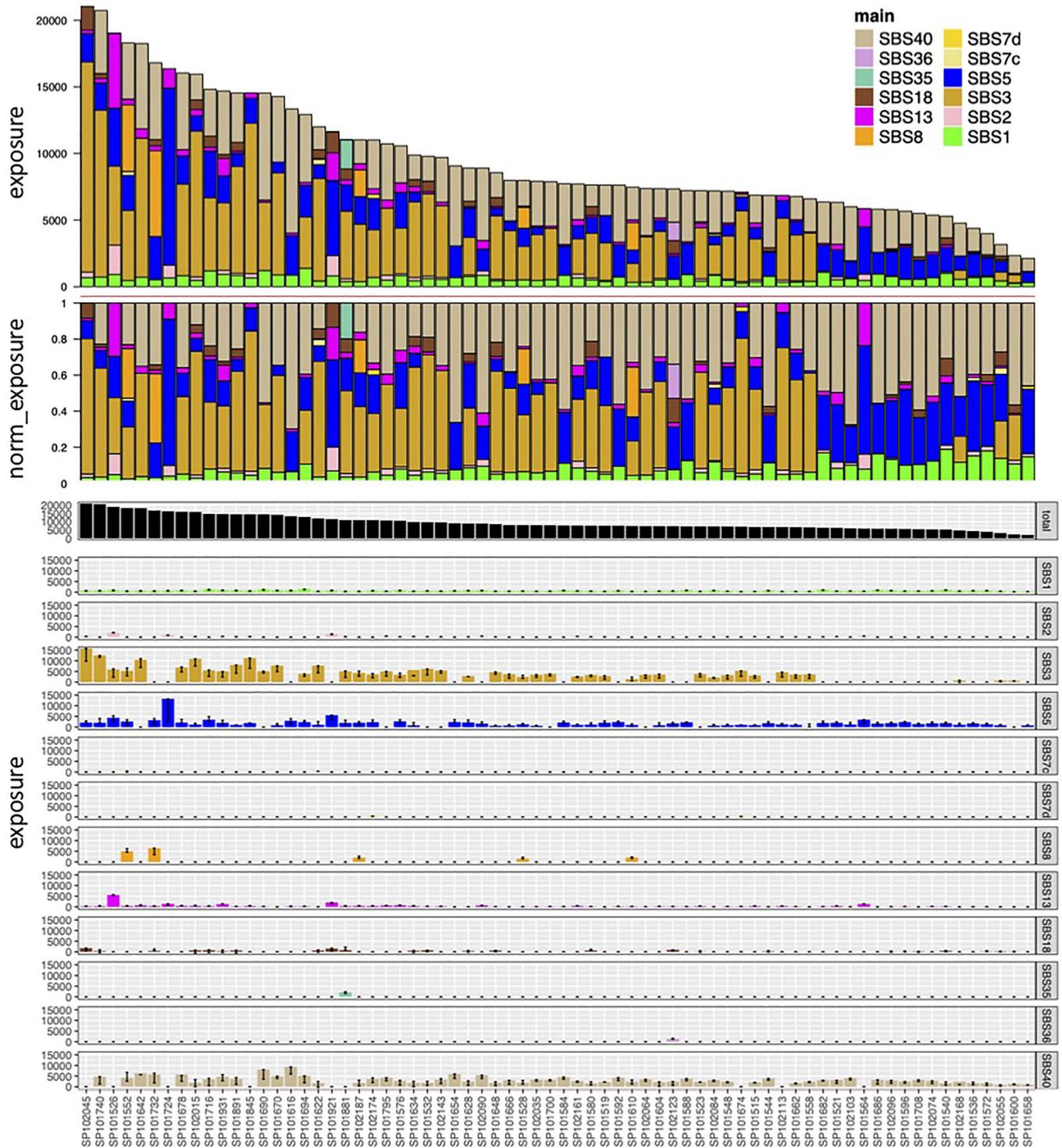
FIGURE 3 Per-sample analysis of an ovarian cancer cohort (same as in Figure 2) with PCAWG SNV signatures. Top panel: absolute exposures; the y-axis displays the number of SNVs explained by the respective mutational signatures. Middle panel: normalized exposures. Bottom panel: absolute exposures with 95% CIs; the y-axis displays the number of SNVs explained by the respective mutational signatures. Samples in all panels are ordered on the x-axis by decreasing SNV mutational load. SNV, single nucleotide variants

3 | RESULTS

YAPSA performs fitting of mutational signatures in a four-step procedure (flow chart at the center of Figure 1A): (a) NNLS

decomposition of the mutational catalogue with the provided signatures; (b) filtering out those signatures which have contributions less than signature-specific cutoffs (provided in the package); (c) rerun an NNLS with only those signatures left after filtering;
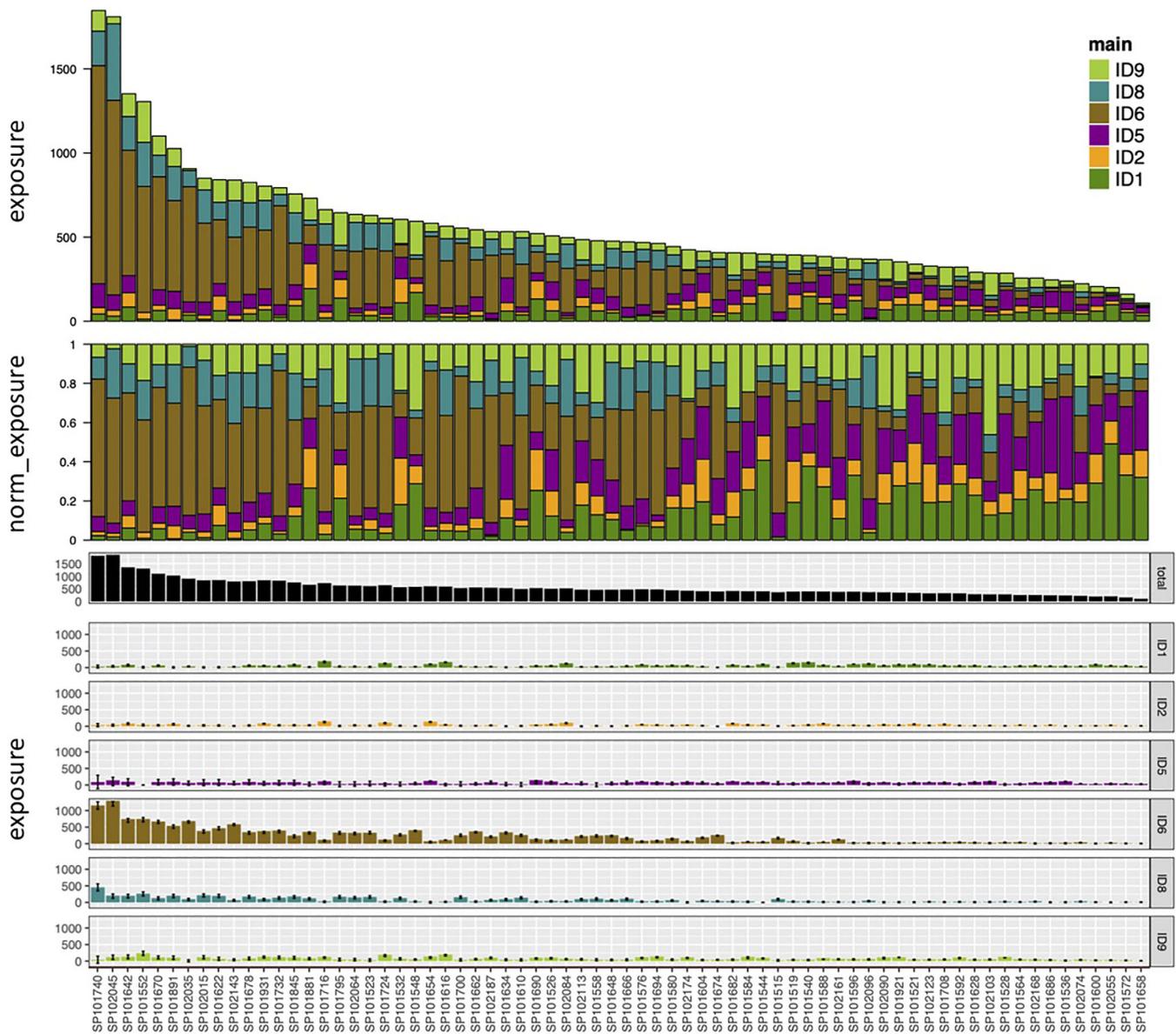
FIGURE 4 Cohort-wide analysis of an ovarian cancer cohort with PCAWG Indel signatures. Top panel: absolute exposures; the y-axis displays the number of Indels explained by the respective mutational signatures. Middle panel: normalized exposures. Bottom panel: absolute exposures with 95% CIs; the y-axis displays the number of SNVs explained by the respective mutational signatures. Samples in all panels are ordered on the x-axis by decreasing Indel mutational load. SNV, single nucleotide variants

and (d) compute CIs. YAPSA furthermore provides functionality to correct for different trinucleotide content and can therefore be used to analyze mutation calls from WES. Several detailed vignettes provide additional information and examples on the usage of YAPSA, starting with https://bioconductor.org/packages/3.12/bioc/vignettes/YAPSA/inst/doc/YAPSA.html. In this work, we demonstrate the application of YAPSA for fitting mutational signatures to a publicly available dataset of ovarian cancer.[29] We show how the identification of specific mutational signatures can serve as biomarker for HRR deficiency and can help predict sensitivity to PARP inhibition. We furthermore show instructive examples of the application of YAPSA in the MASTER precision oncology program.[21,22]

3.1 | Different sets of mutational signatures

YAPSA provides different collections of signatures for mutational signature analysis: (a) COSMIC SNV signatures (referred to as COSMIC V2 signatures at https://cancer.sanger.ac.uk/cosmic/signatures_v2), (b) PCAWG SNV signatures (referred to as COSMIC V3 signatures at https://cancer.sanger.ac.uk/cosmic/signatures_v3), and (c) PCAWG Indel signatures (https://cancer.sanger.ac.uk/cosmic/signatures/ID). The patterns underlying these signatures are stored as data frames in the software package. To unambiguously identify the used signature set, in YAPSA we denominate the COSMIC SNV signatures as AC1 - AC30 (as abbreviation for Alexandrov COSMIC), the PCAWG SNV signature as SBS1 - SBS67, and the PCAWG Indel signatures ID1-ID17. The command
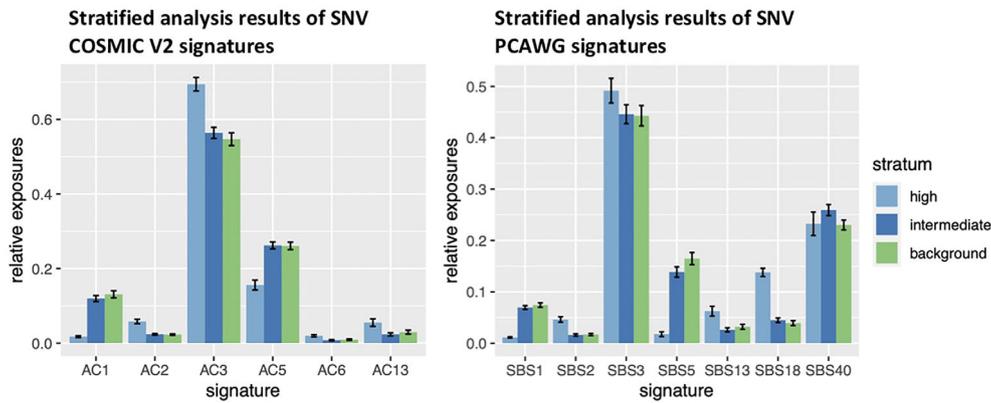
FIGURE 5 Stratified analyses of mutational signatures with enrichment and depletion patterns. The mutational catalog is decomposed into different strata of mutation density (high, intermediate, and background, for details see main text "1.1. Stratification of the mutational catalog"). The ovarian cancer data set was analyzed twice: once using the COSMIC V2 SNV, signatures (left panel) and once using the PCAWG SNV signatures (right panel). The y-axis represents normalized contributions of the different signatures in the respective strata. Some signatures (including the APOBEC-related signatures AC2 and SBS2) are enriched in regions of high mutation density, some (including the aging-related signatures AC1, SBS1, AC5 and SBS5) are depleted in regions of high mutation density. Error bars: SEM. SNV, single nucleotide variants

data (sigs)

loads the COSMIC V2 SNV signatures to the R workspace, whereas

data (sigs_pcawg)

loads the PCAWG SNV and Indel signatures to the R workspace.

Further information and details can be found in the vignettes provided at the Bioconductor website: http://bioconductor.org/packages/3.12/bioc/vignettes/YAPSA/inst/doc/YAPSA.html and http://bioconductor.org/packages/3.12/bioc/vignettes/YAPSA/inst/doc/vignettes_Indel.html

3.2 | Signature-specific cutoffs

To enable highly specific fitting of mutational signatures, YAPSA provides signature-specific cutoffs for all provided signature sets. These are threshold values trained individually for each signature based on the same data from which the respective signature sets had been extracted[1,5] by a modified ROC analysis using the R package ROCR.[26]

In a modified ROC analysis, a cost function is defined by specifying costs punishing for false negative and false positive findings separately. In the implementation of the package ROCR,[26] the cost function is unambiguously defined by the ratio of the cost for a false negative finding divided by the cost for a false positive finding, which in the following we call the $cost_{factor}$:

$$cost_{factor} = \frac{cost\ for\ false\ negative\ findings}{cost\ for\ false\ positive\ findings} \quad (13)$$

The absolute values of the costs for false negative and false positive findings have no effect on the shape of the cost function; this shape and hence the minimum of the cost function depend only on the value of $cost_{factor}$. YAPSA provides sets of optimal signature-specific cutoffs for a range of values of $cost_{factor}$ (for the COSMIC SNV and PCAWG Indel signatures, $cost_{factor}$ was varied in the range [1,... 10], for the PCAWG SNV signatures, $cost_{factor}$ was varied in the range [1,... 15]).

For the different values of $cost_{factor}$, the total number of false attributions was computed (sum of false positive and false negative identifications of signatures). In YAPSA, we chose that $cost_{factor}$ to be optimal for which the total number of false attributions is minimal. Using this criterion, the optimal $cost_{factor}$ for COSMIC SNV signatures was 6, the optimal $cost_{factor}$ for PCAWG SNV signatures was 10, and the optimal $cost_{factor}$ for PCAWG Indel signatures was 3.

Individual ROC analyses were performed for every signature in the different sets of signatures. Supplementary Figure 1 illustrates the cost functions in the ROC analyses for the COSMIC SNV signatures, Supplementary Figure 2 illustrates the cost functions for the PCAWG SNV signatures, and Supplementary Figure 3 the cost functions for the PCAWG Indel signatures for the respective chosen optimal values of $cost_{factor}$. The global minimum of the cost function, indicated by vertical red lines in all these figures, defines the optimal signature-specific cutoff value. These values are displayed in Tables 1-3.

For the clock-like mutational signatures AC1 and AC5 as well as SBS1 and SBS5, optimal cutoffs were set to be zero, as they can be expected to be true positives in all analyses and all cohorts.

In YAPSA, the values for these signature-specific cutoffs are included in the R package. After having loaded the package to the R workspace, these cutoff values are accessible to the user by executing the commands:

data (cutoffs)

for COSMIC SNV mutational signatures or

data (cutoffs_pcawg)

for PCAWG SNV and Indel mutational signatures.

Further information and details can be found in the corresponding vignette: http://bioconductor.org/packages/3.12/bioc/vignettes/YAPSA/inst/doc/vignette_signature_specific_cutoffs.html

Fitting of mutational signatures is only feasible if the data to be analyzed has enough mutations for the NNLS deconvolution to yield reliable results. Especially in the case of Indels, we thus recommend to only use YAPSA for an analysis with PCAWG Indel mutational signatures with WGS data. As the Indel Signature ID15 was not determined through NMF based on WGS data, no signature-specific cutoff could be computed for ID15; this signature is therefore not displayed in Supplementary Figure 3 and in Table 3.

For the ovarian cancer data set,[29] cohort-wide fitting of mutational signatures with YAPSA was performed with all three available sets of mutational signatures with the respective sets of signature-specific cutoffs. The results are show in Figure 2 for the PCAWG SNV signatures, in Figure 4 for the PCAWG Indel signatures and in Supplementary Figure 4 for the COSMIC V2 signatures. In these cohort-wide analyses, signatures associated with aging (AC1 and AC5 in COSMIC V2 SNV signatures as well as SBS1 and SBS5 in PCAWG SNV signatures), APOBEC enzymes (AC2 and AC13 in COSMIC V2 SNV signatures as well as SBS2 and SBS13 in PCAWG SNV signatures), HRR deficiency (AC3 in COSMIC V2, SBS3 in PCAWG SNV, as well as ID6 and ID8 in PCAWG Indel signatures), mismatch repair deficiency (MMR) (AC6 in COSMIC V2 SNV signatures as well as ID1 and ID2 in PCAWG SNV signatures), and exposure to reactive oxygen species (SBS18 in PCAWG SNV signatures) were identified. Furthermore, the analyses detected presence of mutational signatures for which the underlying mutational processes are unknown: SBS40 in PCAWG SNV signatures as well as ID5 and ID9 in PCAWG Indel signatures (Figures 2-3 and Supplementary Figure 4).

In accordance with the design of the signature-specific cutoffs, the signatures identified using the YAPSA algorithm showed very high overlap and consistency with those signatures identified in the original NMF-based signature extraction, cf. Alexandrov et al. 2020.[5]

The ovarian cancer data set[29] was also analyzed at per-sample level (Figure 3 for the PCAWG SNV signatures, Supplementary Figure 5 for the COSMIC V2 SNV signatures and Supplementary Figure 6 for the PCAWG Indel signatures). All signatures identified in the cohort-wide analysis were recovered in the per-sample analysis. In addition to these, several signatures were identified at small frequencies exclusively in the per-sample analysis (Supplementary Information Section "1.1 Cohort-wide and per-sample analyses of mutational signatures"). All of these additional and exclusive signatures together accounted for small fractions of all exposures in the cohort: 22321.38/666947 = 3.35% for the PCAWG SNV signatures, 114 513.2/666947 = 17.17% for the COSMIC V2 SNV signatures and 1708.969/38754 = 4.41% for the PCAWG Indel signatures.

3.3 | Confidence intervals

CIs are computed using the concept of profile likelihood.[27] A detailed vignette about the computation of CIs can be found at http://bioconductor.org/packages/3.12/bioc/vignettes/YAPSA/inst/doc/vignette_confidenceIntervals.html.

For the ovarian cancer data set, CIs for exposures were calculated and displayed for the PCAWG SNV signatures (Figures 2 and 3), the PCAWG Indel signatures (Figure 4 and Supplementary Figure 6) and the COSMIC V2 signatures (Supplementary Figures 4 and 5). Detection of a signature in the mutational catalog of a sample was termed to be "high confidence" if (i) the signature was detected at all and (ii) CI computed for the exposure to this signature in the respective sample excludes zero. As opposed to that, if the CI of the exposure to a signature in a sample included zero, we termed this signature to be detected with "low confidence" in the respective sample.

According to this nomenclature, signatures associated with aging were detected in all samples of the cohort and this calling had high confidence in almost all samples of the cohort: out of 70 samples, 65 (92.86%) and 5 (7.14%) had high and low confidence calls for AC1, respectively, 70 (100%) and 0 (0%) had high and low confidence calls for SBS1, 64 (91.43%) and 6 (8.57%) had high and low confidence calls for AC5, and 50 (71.43%), 16 (22.86%) and 4 (5.71%) had high confidence calls, low confidence calls or no detection for SBS5, respectively.

The fraction of high confidence calls for the signatures associated with APOBEC activity was lower: 58 (82.86%) and 12 (17.14%) samples had high and low confidence calls for AC2, 42 (60.00%), 24 (34.29%) and 4 (5.71%) samples had high confidence calls, low confidence calls or no detection for SBS2, 36 (51.43%), 24 (34.29%) and 10 (14.28%) samples had high confidence calls, low confidence calls or no detection for AC13 and 50 (71.43%), 14 (20.00%) and 6 (8.57%) samples had high confidence calls, low confidence calls or no detection for SBS13, respectively. For some signatures, including AC6 (associated with MMR deficiency), hardly any detection was high confidence: 1 (1.43%), 31 (44.29%) and 38 (54.28%) samples had high confidence calls, low confidence calls or no detection for AC6, respectively.

Of note, the CIs for the exposures to signatures associated with HRR deficiency (SBS3 in PCAWG SNV, AC3 in COSMIC V2 and ID6 and ID8 in PCAWG Indel signatures) revealed high fractions of high confidence calls: 70 (100%) samples had high confidence calls for AC3, 65 (92.86%) and 5 (7.14%) samples had high and low confidence calls for SBS3, 64 (91.43%) and 6 (8.57%) samples had high and low confidence calls for ID6, and 52 (74.28%), 17 (24.29%) and 1 (1.43%) samples had high confidence calls, low confidence calls or no detection for ID8, respectively. The presence of this mutational mechanism in this cohort can thus be detected with high confidence in this ovarian cancer cohort.

3.4 | Application of YAPSA to WES data

As WGS and WES differ in the frequency of occurrence of different k-mers, these differences have to be corrected for. In YAPSA, this is performed with the function `normalizeMotifs_otherRownames()`.

TABLE 1 Optimal absolute signature-specific cutoff values for COSMIC SNV mutational signatures in WGS data for a $cost_{factor}$ of 6

Cost factor	AC1	AC2	AC3	AC4	AC5	AC6	AC7	AC8	AC9
6	0	0.010459	0.081941	0.017540	0	0.001549	0.040133	0.242755	0.115171
	AC10	AC11	AC12	AC13	AC14	AC15	AC16	AC17	AC18
6	0.010084	0.099249	0.2106201	0.007877	0.144306	0.037960	0.3674349	0.002648	0.332539
	AC19	AC20	AC21	AC22	AC23	AC24	AC25	AC26	AC27
6	0.115645	0.123503	0.164026	0.031022	0.0333866	0.032402	0.016119	0.093352	0.009320
	AC28	AC29	AC30						
6	0.056164	0.059362	0.059153						

Of note, WES can be performed with different target capture kits. As these cover different genomic regions, the correction factors may vary between the different target capture kits. For target capture-specific correction, that is, for a given target capture kit A, correction factors q_X^{WGS,WES_A} for all features have to be computed, and YAPSA allows for this specificity. As detailed below, correction factors for nine different target capture kits and one correction factor directly derived from the gene model GENCODE 19 applied to the human reference genome hs37d5 are stored in YAPSA. The available correction factors can be loaded to the R workspace by the following command:

    data (targetCapture_cor_factors)

Using the command

    names (targetCapture_cor_factors)

the user can see that correction factors for the capture kits

- Agilent4withUTRs
- Agilent4withoutUTRs
- Agilent5withUTRs
- Agilent5withoutUTRs
- Agilent6withUTRs
- Agilent6withoutUTRs
- Agilent7withoutUTRs
- AgilentSureSelectAllExon
- IlluminaNexteraExome

are provided. When correcting, for example, for the triplet content in the target capture kit *Agilent SureSelect all exon*, the function `normalizeMotifs_otherRownames()`

    targetCapture <- " AgilentSureSelectAllExon "
    cor_list <- targetCapture_cor_factors [[targetCapture]]
    corrected_catalog_df <-
    normalizeMotifs_otherRownames
     (exome_mutCatRaw_df, cor_list$rel_cor)

can be called as follows:

More information and a detailed example of application of YAPSA to WES data can be found in a dedicated vignette:

https://www.bioconductor.org/packages/release/bioc/vignettes/YAPSA/inst/doc/vignette_exomes.html

3.5 | Stratification of the mutational catalog

To further characterize the properties of mutational processes, SNVs can be assigned into different categories (termed strata in the following), for which signature enrichment or depletion patterns can be computed. Performing separate analyses of mutational signatures on the strata individually can be error-prone: As the statistical power of a stratum is always lower than the power of the entire set of mutations, separate analyses are prone to yield signature calls in individual strata which include signatures that have not been present in the analysis of the entire set of mutations. Such errors can be avoided by a constrained stratified analysis.

The strata have to be exclusive, that is, every SNV must be in exactly one stratum. Examples for strata are genomic regions with low, intermediate or high mutation density, genomic regions with early or late replication timing, and clonal or subclonal mutations. This is also of particular interest when studying localized mutational processes (cf. Maura et al.[12]). YAPSA provides the function SMC (Stratified analysis of mutational signatures) to solve the stratified optimization problem (Methods Section "3.5 Stratification of the mutational catalogue"). Stratified analyses of mutational signatures are also covered in a dedicated vignette: http://bioconductor.org/packages/3.12/bioc/vignettes/YAPSA/inst/doc/vignette_stratifiedAnalysis.html

For the ovarian cancer data set, we performed a stratification of all SNVs by mutation density. SNVs with intermutation distance ≤1 kbp were assigned to a stratum of high mutation density, those with 1 kbp < intermutation distance ≤100 kbp were assigned to a stratum of intermediate mutation density, and those with intermutation density > 100 kbp were assigned to a stratum called background.

In both analyses with COSMIC V2 and PCAWG SNV signatures, the signatures associated with APOBEC enzyme activity, AC2 and SBS2, showed high enrichment in the stratum of high mutation density (AC2, Kruskal-Wallis [KW] test, multiple testing correction

TABLE 2 Optimal signature-specific cutoff values for PCAWG SNV mutational signatures for a $cost_{factor}$ of 10. Cutoffs are valid for the analysis of both WGS and WES data. For the clock-like mutational signatures SBS1 and SBS5, optimal cutoffs are set to be zero, as they are true positives in all analyses and all cohorts

Cost factor	SBS1	SBS2	SBS3	SBS4	SBS5	SBS6	SBS7a	SBS7b	SBS7c
10	0	0.013323	0.001706	0.081066	0	0.21938	0.107568	0.097096	0.010685
	SBS7d	SBS8	SBS9	SBS10a	SBS10b	SBS11	SBS12	SBS13	SBS14
10	0.017561	0.25758	0.123529	0.085963	0.056244	0.124596	0.173822	0.011157	0.082277
	SBS15	SBS16	SBS17a	SBS17b	SBS18	SBS19	SBS20	SBS21	SBS22
10	0.062881	0.328003	0.02977	0.031387	0.005704	0.193317	0.11269	0.174782	0.129109
	SBS23	SBS24	SBS25	SBS26	SBS28	SBS29	SBS30	SBS31	SBS32
10	0.137645	0.158176	0.307735	0.246075	0.131058	0.109774	0.147923	0.164218	0.243039
	SBS33	SBS34	SBS35	SBS36	SBS37	SBS38	SBS39	SBS40	SBS41
10	0.169249	0.170554	0.128734	0.155968	0.235212	0.101551	0.374672	0	0.163599
	SBS42	SBS44							
10	0.138322	0.158954							

TABLE 3 Optimal signature specific cutoff values for PCAWG Indel mutational signatures for a $cost_{factor}$ of three. Cutoffs are valid for the analysis of WGS data. For signatures ID1 and ID2, optimal cutoffs are set to be zero, as they are true positives in all analyses and all cohorts

Cost factor	ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	ID9
3	0	0	0.068509	0.159711	0.021578	0.064558	0.290663	0.049468	0.069298
	ID10	ID11	ID12	ID13	ID14	ID16	ID17		
3	0.179891	0.049493	0.097596	0.127275	0.077038	0.317817	0.158396		

according to Benjamini and Hochberg: p_{KW} = 4.11*10^{-9}, SBS2: p_{KW} = 6.16*10^{-8}), whereas signature SBS13 showed a still significant, but less pronounced enrichment (p_{KW} = 3.88*10^{-2}) and signature AC13 showed only a trend (p_{KW} = 3.43*10^{-1}). Complementarily, the aging signatures were depleted in the stratum of mutations with high mutation density (AC1: p_{KW} = 6.12*10^{-26}, AC5: p_{KW} = 1.94*10^{-8}, SBS1: p_{KW} = 7.37*10^{-29}, SBS5: p_{KW} = 3.41*10^{-22}).

Enrichment and depletion patterns for signatures associated with HRR deficiency showed an enriched in the stratum of SNVs with high mutation density in the analysis with COSMIC V2 signatures (AC3: p_{KW} = 2.19*10^{-7}) and a trend in the analysis with PCAWG SNV signatures (SBS3: p_{KW} = 1.54*10^{-1}).

3.6 | Mutational signatures in precision oncology

YAPSA has been used for analysis of mutational signatures in the Molecular Tumor Board (MTB) of the MASTER program[21,22] since 2015. Here, we present two cases that were analyzed in this framework.

Case1 (Figure 6A-C) was a woman with uterine leiomyosarcoma and a pathogenic frameshift deletion in *BRCA2* (Figure 6B). A tumor specimen was subjected to WGS in October 2017. A total of 10 866 somatic SNVs and 926 somatic Indels were identified. This case was characterized by the HRR defect signatures AC3 (explaining 4108 or 37.8% of the somatic SNVs), SBS3 (explaining 3557 or 32.7% of the somatic SNVs) and ID6 (explaining 392 or 42.4% of the somatic Indels) when fitting mutational signatures with YAPSA (Figure 6A). All these signatures were detected with high confidence (confidence

interval excluding zero, Supplementary Table 1). This result matched the finding of increased genomic instability as evidenced by an LOH-HRD score[33] of 23 and 20 LSTs[36] detected in this triploid sample (Figure 6C). These observations were in line with a germline loss-of-function mutation in *BRCA2*. PARP inhibition alone or in combination may be a therapeutic option for this patient.

Case2 (Figure 6D-E) was a woman with a neuroendocrine neoplasm, which was subjected to WGS in January 2020. Neither germline nor somatic mutations in genes associated with the HRR pathway were identified (Figure 6E). Furthermore, fitting of mutational signatures with YAPSA did not identify AC3, SBS3, ID6, or ID8 (Figure 6D, Supplementary Table 2). LOH-HRD[33] score and the number of LSTs[36] were zero (even though the tetraploid genome did exhibit whole chromosome sCNAs, but these are excluded in the computation of LOH-HRD and LSTs, Figure 6F). No arguments for the use of PARP inhibitors were found in this patient.

These two cases illustrate the congruence between mutational signatures and the quantification of genomic instability by LOH-HRD scores and the number of LSTs. Furthermore, the signature analyses with the different sets of mutational signatures are consistent between each other with respect to the detection of HRR-associated signatures.

4 | DISCUSSION

YAPSA is a user-friendly R/Bioconductor package for fitting mutational signatures using SNV and Indel signatures from the COSMIC or PCAWG signature sets. It uses optimal signature-specific cutoffs to
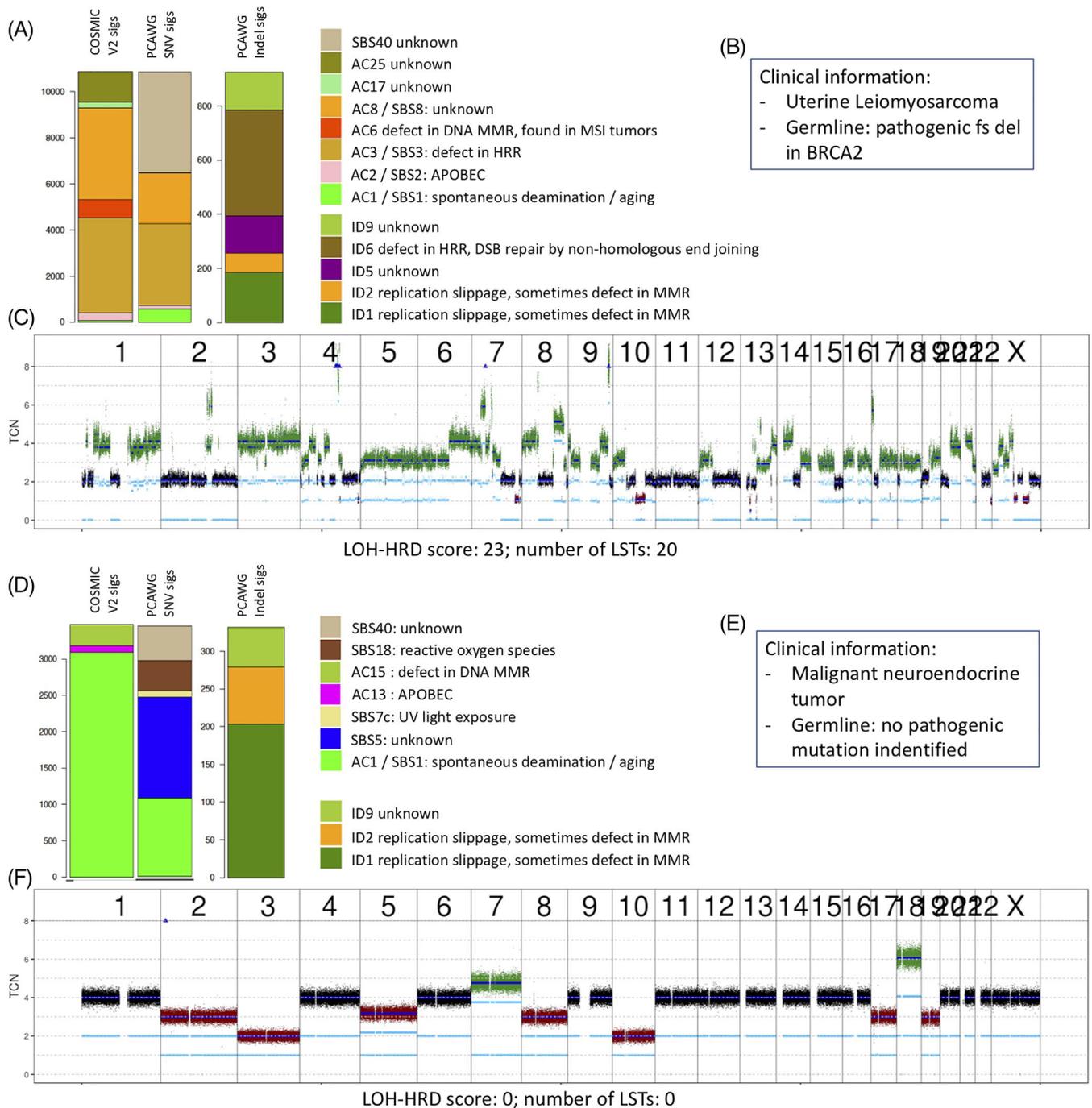
FIGURE 6 Exemplary cases from the NCT/DKTK MTB. A, B, and C: *Case1*, a woman with uterine leiomyosarcoma and a germline *BRCA2* frameshift deletion; D, E, and F,: *Case2*, a woman with a neuroendocrine neoplasm without mutations in genes of the HRR pathway. A, and D, display the results of the fitting of mutational signatures with YAPSA; B, and E, highlight clinical and germline information; C, and F, show copy number plots. Note that whole chromosome events contribute neither to the LOH-HRD score nor to the number of LSTs. HRR, homologous recombination repair; LSTs, large scale state transitions; LOH-HRD, loss-of-heterozygosity-homologous-recombination-deficiency; YAPSA, yet another package for signature analysis

reduce false positive calls and provides CIs as a measure of the uncertainty of determined signature contributions. A functionality for stratified analyses of mutational signatures enables the investigation of enrichment and depletion patterns in subsets of mutations. Together with additional functionalities and various visualization capabilities, this makes YAPSA a comprehensive package to analyze activities of

mutational processes in cancer cohorts of any size, including individual patient samples.

When applied to an ovarian cancer dataset,[5,29] this functionality recovered mutational mechanisms established to be active in this cancer type, reflected by aging signatures, signatures associated with APOBEC enzyme activity, and signatures associated with HRR

defects.[1,5,29] Using the concept of signature-specific cutoffs and thereby increasing specificity in signature analysis, YAPSA lead to the detection of these mutational mechanisms consistently across different sets of mutational signatures: the COSMIC V2 and PCAWG SNV signatures as well as PCAWG Indel signature sets. In addition to the sole detection of signatures and the computation of the respective exposures, the possibility to assess the confidence of the detection is of particular use when using mutational signatures for treatment recommendation.

When applied to a whole cohort, fitting of mutational signatures can be performed cohort-wide or at per-sample level. The two methods are complementary. A cohort-wide analysis, on one hand, identifies mutational signatures and reveals mutational mechanisms active in the whole cohort with high specificity. The per-sample analysis, on the other hand, is more sensitive and can capture additional processes active in a single sample, which may have been diluted in the cohort-wide analysis. However, the per-sample analysis is less robust to technical noise, is less specific, and may lead to false positive calls. When applied to the example of the ovarian cancer cohort,[29] the per-sample analysis recovered all signatures identified in the cohort-wide analysis. Furthermore, in this data set, the contribution of signatures identified exclusively in the per-sample analysis and not in the cohort-wide analysis was rather small (3.35% for the PCAWG SNV signatures, 17.17% for the COSMIC V2 SNV signatures and 4.41% for the PCAWG Indel signatures). These figures capture both the amount of technical noise and the biological heterogeneity of the samples. A detailed computation of the respective contributions of these two sources of variation may only be possible by large scale benchmarks in future research. Of note, in a setting where the mutational load per sample is very low, per-sample analyses may not be feasible and cohort-wide analyses may be the only meaningful alternative.

As shown previously, some mutational signatures may be used as biomarkers.[12,37,38] This is particularly well established for signatures associated with deficiency in HRR,[39,40] that is, signatures AC3, SBS3 and/or ID6 and ID8. Deficiency in HRR might represent a suitable target for therapeutic intervention using agents, either alone or in combination, that are preferentially toxic to HRR-incompetent cells, such as PARP inhibitors, platinum derivatives, or trabectedin.[41] Primarily in BRCA1/2-deficient epithelial cancers, notably breast, ovarian, and prostate cancer, the concept of "BRCAness" was introduced to characterize an endo-phenotype with various imprints of the DNA repair defect on the genome, including genomic scars and genomic instability.[33,36,42,43] Especially in ovarian cancer, BRCAness is frequent and reflects an established synthetic lethal relationship with pharmacologic PARP inhibition[44,45] - leading to the latter class of drugs being approved for treatment in this entity.[46-49] The detection of the HRR-associated signatures AC3, SBS3, ID6 and ID8 in the ovarian cancer cohort in this work is in line with this.

Fitting mutational signatures, when used as a detection tool for HRR-associated signatures (AC3, SBS3, ID6 and ID8), represents a biomarker which is complementary to the detection of genomic scars and/or causative mutations in genes of the HRR pathway. Using this concept, we and others have identified BRCAness in various other entities, for example, osteosarcoma,[50,51] leiomyosarcoma,[52] in which efficacy of PARP1 inhibition has been shown in vitro[52] and in preclinical models,[53] or chordoma.[4] Furthermore, data from a phase 1b trial of olaparib and trabectedin in unselected patients with relapsed bone and soft-tissue sarcoma suggest that this treatment might be effective in subgroups of these entities.[54] Additional clinical benefit from sensitive and specific biomarkers of HRR-deficiency may potentially be obtained by using immune checkpoint blockade in HRR-deficient tumors and/or combining it with PARP inhibition.[55-57]

HRR-associated signatures have also been detected in samples from cancer patients who had been exposed to ionizing radiation, exemplified in chordomas[4] as well as in meningiomas.[58,59] Even though the imprint of ionizing radiation in cancer is not well understood,[60] using YAPSA, comparison of a cohort of low-dose radiation induced meningiomas with a cohort of sporadic meningiomas revealed that in sporadic meningiomas, BRCAness was associated with potential causing mutations in genes associated with homologous recombination repair, whereas in low-dose radiation induced meningiomas, this was not the case.[59] In addition, an enrichment of the HRR-associated signature AC3 in the vicinity of (potentially radiation-induced) breakpoints of structural variants was found only in low-dose radiation induced meningiomas, arguing in favor of AC3 being linked to an exhausted capacity of HRR at the loci of radiation-induced breakpoints.[59] This demonstrates how the functionality of YAPSA to perform stratified analyses of mutational signatures can be used to reveal biologically relevant enrichment and depletion patterns.

Highly precise determination of mutational signature contributions is particularly useful in precision oncology programs that employ broad genomic profiling to identify targetable lesions in cancer patients who have exhausted standard therapy options.[13-20] As mentioned earlier, the MASTER program is a multicenter registry trial for patients with rare tumors and for younger adults with advanced cancer across all entities.[21,22] With its functionality of signature-specific cutoffs and the possibility to indicate confidence measures for signature calls, YAPSA has proved to be a valuable tool for fitting mutational signatures in this precision oncology context.[4,52,61-63] To date, more than 2000 patients have been analyzed with YAPSA.

In summary, YAPSA is a tool for fitting mutational signatures with various unique features: signature-specific cutoffs reduce false positive calls, CIs are a measure of uncertainty in the detection of signatures in a patient or cohort, and stratified analyses of mutational signatures yield enrichment and depletion patterns of signature exposures in subsets of the detected mutations, thereby enabling the interpretation of underlying mutational mechanisms.

ACKNOWLEDGEMENTS

DATA AVAILABILITY STATEMENT

External data: For the data used to generate Figure 1, we refer to the Supplementary Information section "3 External data". In this work, the functionality of the software package YAPSA is also demonstrated on an ovarian cancer data set29. Data was downloaded on the 4th of June 2020 from https://dcc.icgc.org/api/v1/download?fn=/current/Projects/OV-AU/simple_somatic_mutation.open.OV-AU.tsv.gz. Only those samples with more than 25 SNVs and more than 20 Indels were kept for analysis. The whole analysis used to generate Figures 2 - 5 and Supplementary Figures 4 - 6 as well as all numeric values for this cohort are provided in the supplementary files Code_for_figure_generation.Rmd (R markdown) and Code_for_figure_generation.html (compiled report). Processing of samples from the precision oncology program MASTER: Two samples presented in this work underwent paired-end WGS (2 × 151 bp) on a HiSeq X instrument (Illumina, San Diego, California) in the framework of the MASTER program. Library preparation was performed with the TruSeq Nano Library Preparation Kit (Illumina). Alignment with BWA mem, small variant calling and calling of somatic copy number aberrations (sCNAs) with ACEseq30 were performed as described earlier4,31,32. Somatic SNV (case1_somaticSnvs.vcf, case2_somaticSnvs.vcf) and somatic Indel (case1_somaticIndels.vcf, case2_somaticIndels.vcf) variant calls required for reproducibility of the analysis of mutational signatures as well as segment information obtained from sCNA calling (case1_segments.csv, case2_segments.csv) required for the reproducibility of the computation of genomic instability (cf. below) are provided as supplementary files.

ORCID

Daniel Hübschmann https://orcid.org/0000-0002-6041-7049
Carolin Andresen https://orcid.org/0000-0002-8960-7719
Christoph E. Heilig https://orcid.org/0000-0001-8869-1421

REFERENCES

1. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500:415-421. https://doi.org/10.1038/nature12477.
2. Rubanova Y, Shi R, Harrigan CF, et al. Reconstructing evolutionary trajectories of mutation signature activities in cancer using TrackSig. *Nat Commun*. 2020;11(1):731. https://doi.org/10.1038/s41467-020-14352-7.
3. Van Hoeck A, Tjoonk NH, van Boxtel R, Cuppen E. Portrait of a cancer: mutational signature analyses for cancer diagnostics. *BMC Cancer*. 2019;19(1):457. https://doi.org/10.1186/s12885-019-5677-2.
4. Gröschel S, Hübschmann D, Raimondi F, et al. Defective homologous recombination DNA repair as therapeutic target in advanced chordoma. *Nat Commun*. 2019;10(1):1635. https://doi.org/10.1038/s41467-019-09633-9.
5. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. *Nature*. 2020;578(7793):94-101. https://doi.org/10.1038/s41586-020-1943-3.
6. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol*. 2016;17(1):31. https://doi.org/10.1186/s13059-016-0893-4.
7. Shinde J, Bayard Q, Imbeaud S, et al. Palimpsest: an R package for studying mutational and structural variant signatures along clonal evolution in cancer. *Bioinformatics*. 2018;34(19):3380-3381. https://doi.org/10.1093/bioinformatics/bty388.
8. Huang PJ, Chiu LY, Lee CC, et al. MSignatureDB: a database for deciphering mutational signatures in human cancers. *Nucleic Acids Res*. 2018;46(D1):D964-D970. https://doi.org/10.1093/nar/gkx1133.
9. Huang X, Wojtowicz D, Przytycka TM. Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics*. 2018;34:330-337. https://doi.org/10.1093/bioinformatics/btx604.
10. Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med*. 2018;10(1):33. https://doi.org/10.1186/s13073-018-0539-0.
11. Schumann F, Blanc E, Messerschmidt C, Blankenstein T, Busse A, Beule D. SigsPack, a package for cancer mutational signatures. *BMC Bioinformatics*. 2019;20(1):450. https://doi.org/10.1186/s12859-019-3043-7.
12. Maura F, Degasperi A, Nadeu F, et al. A practical guide for mutational signature analysis in hematological malignancies. *Nat Commun*. 2019;10(1):2969. https://doi.org/10.1038/s41467-019-11037-8.
13. Willemsen AECAB, Krausz S, Ligtenberg MJL, et al. Molecular tumour boards and molecular diagnostics for patients with cancer in The Netherlands: experiences, challenges, and aspirations. *Br J Cancer*. 2019;121(1):34-36. https://doi.org/10.1038/s41416-019-0489-3.
14. van der Velden DL, Hoes LR, van der Wijngaart H, et al. The drug rediscovery protocol facilitates the expanded use of existing anticancer drugs. *Nature*. 2019;574(7776):127-131. https://doi.org/10.1038/s41586-019-1600-x.
15. Priestley P, Baber J, Lolkema MP, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature*. 2019;575(7781):210-216. https://doi.org/10.1038/s41586-019-1689-y.
16. Le Tourneau C, Delord J-P, Gonçalves A, et al. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. *Lancet Oncol*. 2015;16(13):1324-1334. https://doi.org/10.1016/S1470-2045(15)00188-6.
17. Schwaederle M, Parker BA, Schwab RB, et al. Precision oncology: the UC san Diego Moores cancer center PREDICT experience. *Mol Cancer Ther*. 2016;15(4):743-752. https://doi.org/10.1158/1535-7163.MCT-15-0795.
18. Roychowdhury S, Iyer MK, Robinson DR, et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci Transl Med*. 2011;3(111):111ra121. https://doi.org/10.1126/scitranslmed.3003161.
19. Cheng DT, Mitchell TN, Zehir A, et al. Memorial Sloan Kettering-integrated mutation profiling of actionable cancer targets (MSK-

IMPACT). *J Mol Diagnostics*. 2015;17(3):251-264. https://doi.org/10.1016/j.jmoldx.2014.12.006.

20. Cheng DT, Prasad M, Chekaluk Y, et al. Comprehensive detection of germline variants by MSK-IMPACT, a clinical diagnostic platform for solid tumor molecular oncology and concurrent cancer predisposition testing. *BMC Med Genomics*. 2017;10(1):33. https://doi.org/10.1186/s12920-017-0271-4.

21. Horak P, Klink B, Heining C, et al. Precision oncology based on omics data: the NCT Heidelberg experience. *Int J Cancer*. 2017;141(5):877-886. https://doi.org/10.1002/ijc.30828.

22. Horak P, Fröhling S, Glimm H. Integrating next-generation sequencing into clinical oncology: strategies, promises and pitfalls. *ESMO Open*. 2016;1(5):e000094. https://doi.org/10.1136/esmoopen-2016-000094.

23. Mullen KM, Stokkum IHM. *The Lawson-Hanson Algorithm for Non-negative Least Squares (NNLS)*. CRAN; 2012. https://cran.r-project.org/web/packages/nnls.

24. Gehring JS, Fischer B, Lawrence M, Huber W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics*. 2015;31(22):3673-3675. https://doi.org/10.1093/bioinformatics/btv408.

25. Pagès H, Aboyoun P, Gentleman R, DebRoy S. *Biostrings: String objects representing biological sequences, and matching algorithms* 2016. https://bioconductor.org/packages/release/bioc/html/Biostrings.html.

26. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics*. 2005;21(20):3940-3941. https://doi.org/10.1093/bioinformatics/bti623.

27. Raue A, Kreutz C, Maiwald T, et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*. 2009;25(15):1923-1929. https://doi.org/10.1093/bioinformatics/btp358.

28. Borchers HW. *pracma*. 2018.

29. Patch A-M, Christie EL, Etemadmoghadam D, et al. Whole–genome characterization of chemoresistant ovarian cancer. *Nature*. 2015;521(7553):489-494. https://doi.org/10.1038/nature14410.

30. Kleinheinz K, Bludau I, Huebschmann D, et al. *ACEseq - allele specific copy number estimation from whole genome sequencing*. bioRxiv. 2017. http://biorxiv.org/content/early/2017/10/29/210807.abstract.

31. Reisinger E, Genthner L, Kerssemakers J, et al. OTP: an automated system for managing and processing NGS data. *J Biotechnol*. 2017;261:53-62. https://doi.org/10.1016/j.jbiotec.2017.08.006.

32. Jones DTW, Jäger N, Kool M, et al. Dissecting the genomic complexity underlying medulloblastoma. *Nature*. 2012;488(7409):100-105. https://doi.org/10.1038/nature11284.

33. Abkevich V, Timms KM, Hennessy BT, et al. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br J Cancer*. 2012;107(10):1776-1782. https://doi.org/10.1038/bjc.2012.451.

34. Wilcoxen KM, Becker M, Neff C, et al. Use of homologous recombination deficiency (HRD) score to enrich for niraparib sensitive high grade ovarian tumors. *J Clin Oncol*. 2015;33(15_suppl):5532. https://doi.org/10.1200/jco.2015.33.15_suppl.5532.

35. Telli ML, Timms KM, Reid J, et al. Homologous recombination deficiency (hrd) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple-negative breast cancer. *Clin Cancer Res*. 2016;22(15):3764-3773. https://doi.org/10.1158/1078-0432.CCR-15-2477.

36. Popova T, Manié E, Rieunier G, et al. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res*. 2012;72(21):5454-5462. https://doi.org/10.1158/0008-5472.CAN-12-1470.

37. Davies H, Morganella S, Purdie CA, et al. Whole-genome sequencing reveals breast cancers with mismatch repair deficiency. *Cancer Res*. 2017;77(18):4755-4762. https://doi.org/10.1158/0008-5472.CAN-17-1083.

38. Ma J, Setton J, Lee NY, Riaz N, Powell SN. The therapeutic significance of mutational signatures from DNA repair deficiency in cancer. *Nat Commun*. 2018;9(1):1-12. https://doi.org/10.1038/s41467-018-05228-y.

39. Davies H, Glodzik D, Morganella S, et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat Med*. 2016;23:525. https://doi.org/10.1038/nm.4292.

40. Gröschel S, Bommer M, Hutter B, et al. Integration of genomics and histology revises diagnosis and enables effective therapy of refractory cancer of unknown primary with PDL1 amplification. *Mol Case Stud*. 2016;2(6):mcs.a001180. https://doi.org/10.1101/mcs.a001180.

41. Lord CJ, Ashworth A. BRCAness revisited. *Nat Rev Cancer*. 2016;16(2):110-120. https://doi.org/10.1038/nrc.2015.21.

42. Fong PC, Boss DS, Yap TA, et al. Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. *N Engl J Med*. 2009;361(2):123-134. https://doi.org/10.1056/NEJMoa0900212.

43. Mateo J, Carreira S, Sandhu S, et al. DNA-repair defects and Olaparib in metastatic prostate cancer. *N Engl J Med*. 2015;373(18):1697-1708. https://doi.org/10.1056/NEJMoa1506859.

44. O'Neil NJ, Bailey ML, Hieter P. Synthetic lethality and cancer. *Nat Rev Genet*. 2017;18(10):613-623. https://doi.org/10.1038/nrg.2017.47.

45. Lord CJ, Ashworth A. PARP inhibitors: synthetic lethality in the clinic. *Science*. 2017;355(6330):1152-1158. https://doi.org/10.1126/science.aam7344.

46. Tucker H, Charles Z, Robertson J, Adam J. NICE guidance on olaparib for maintenance treatment of patients with relapsed, platinum-sensitive, BRCA mutation-positive ovarian cancer. *Lancet Oncol*. 2016;17(3):277-278. https://doi.org/10.1016/S1470-2045(16)00062-0.

47. Ray-Coquard I, Pautier P, Pignata S, et al. Olaparib plus Bevacizumab as first-line maintenance in ovarian cancer. *N Engl J Med*. 2019;381(25):2416-2428. https://doi.org/10.1056/NEJMoa1911361.

48. Ledermann J, Harter P, Gourley C, et al. Olaparib maintenance therapy in patients with platinum-sensitive relapsed serous ovarian cancer: a preplanned retrospective analysis of outcomes by BRCA status in a randomised phase 2 trial. *Lancet Oncol*. 2014;15(8):852-861. https://doi.org/10.1016/S1470-2045(14)70228-1.

49. Lheureux S, Lai Z, Dougherty BA, et al. Long-term responders on Olaparib maintenance in high-grade serous ovarian cancer: clinical and molecular characterization. *Clin Cancer Res*. 2017;23(15):4086-4094. https://doi.org/10.1158/1078-0432.CCR-16-2615.

50. Kovac M, Blattmann C, Ribi S, et al. Exome sequencing of osteosarcoma reveals mutation signatures reminiscent of BRCA deficiency. *Nat Commun*. 2015;6:8940. https://doi.org/10.1038/ncomms9940.

51. Engert F, Kovac M, Baumhoer D, Nathrath M, Fulda S. Osteosarcoma cells with genetic signatures of BRCAness are susceptible to the PARP inhibitor talazoparib alone or in combination with chemotherapeutics. *Oncotarget*. 2017;8(30):48794-48806. https://doi.org/10.18632/oncotarget.10720.

52. Chudasama P, Mughal SS, Sanders MA, et al. Integrative genomic and transcriptomic analysis of leiomyosarcoma. *Nat Commun*. 2018;9(1):144. https://doi.org/10.1038/s41467-017-02602-0.

53. Pignochino Y, Capozzi F, D'Ambrosio L, et al. PARP1 expression drives the synergistic antitumor activity of trabectedin and PARP1 inhibitors in sarcoma preclinical models. *Mol Cancer*. 2017;16(1):86. https://doi.org/10.1186/s12943-017-0652-5.

54. Grignani G, D'Ambrosio L, Pignochino Y, et al. A phase 1b trial with the combination of trabectedin and olaparib in relapsed patients (pts) with advanced and unresectable bone and soft tissue sarcomas (BSTS): an Italian sarcoma group (ISG) study. *J Clin Oncol*. 2016;34(15_suppl):11018. https://doi.org/10.1200/JCO.2016.34.15_suppl.11018.

55. Mathios D, Ruzevick J, Jackson CM, et al. PD-1, PD-L1, PD-L2 expression in the chordoma microenvironment. *J Neurooncol*. 2015;121(2):251-259. https://doi.org/10.1007/s11060-014-1637-5.

56. Nolan E, Savas P, Policheni AN, et al. Combined immune checkpoint blockade as a therapeutic strategy for BRCA1-mutated breast cancer. *Sci Transl Med*. 2017;9(393):eaal4922. https://doi.org/10.1126/scitranslmed.aal4922.

57. Pantelidou C, Sonzogni O, Taveira MDO, et al. Parp inhibitor efficacy depends on CD8+ T-cell recruitment via intratumoral sting pathway activation in brca-deficient models of triple-negative breast cancer. *Cancer Discov*. 2019;9(6):722-737. https://doi.org/10.1158/2159-8290.CD-18-1218.

58. Sahm F, Toprak UH, Hübschmann D, et al. Meningiomas induced by low-dose radiation carry structural variants of NF2 and a distinct mutational signature. *Acta Neuropathol*. 2017;134(1):155-158. https://doi.org/10.1007/s00401-017-1715-9.

59. Paramasivam N, Hübschmann D, Toprak UH, et al. Mutational patterns and regulatory networks in epigenetic subgroups of meningioma. *Acta Neuropathol*. 2019;138(2):295-308. https://doi.org/10.1007/s00401-019-02008-w.

60. Behjati S, Gundem G, Wedge DC, et al. Mutational signatures of ionizing radiation in second malignancies. *Nat Commun*. 2016;7:12605. https://doi.org/10.1038/ncomms12605.

61. Dieter SM, Heining C, Agaimy A, et al. Mutant KIT as imatinib-sensitive target in metastatic sinonasal carcinoma. *Ann Oncol*. 2017;28(1):142-148. https://doi.org/10.1093/annonc/mdw446.

62. Gröschel S, Bommer M, Hutter B, et al. Integration of genomics and histology revises diagnosis and enables effective therapy of refractory cancer of unknown primary with PDL1 amplification. *Cold Spring Harb Mol Case Stud*. 2016;2(6):a001180.

63. Horak P, Weischenfeldt J, von Amsberg G, et al. Response to olaparib in a PALB2 germline mutated prostate cancer and genetic events associated with resistance. *Mol Case Stud*. 2019;5(2):a003657. https://doi.org/10.1101/mcs.a003657.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.