Accurate and sensitive mutational signature analysis with MuSiCal

Hu Jin[1†], Doga C. Gulhan[1†], Daniel Ben-Isvy[1], David Geng[1], Viktor Ljungstrom[1] and Peter J. Park[1*]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, 02115, MA, USA.

*Corresponding author(s). E-mail(s): peter_park@hms.harvard.edu;
†These authors contributed equally to this work.

Abstract

Mutational signature analysis is a recent computational approach for interpreting somatic mutations in the genome. Its application to cancer data has enhanced our understanding of mutational forces driving tumorigenesis and demonstrated its potential to inform prognosis and treatment decisions. However, methodological challenges remain for discovering new signatures and assigning proper weights to existing signatures, thereby hindering broader clinical applications. We present MuSiCal (Mutational Signature Calculator), a rigorous analytical framework with novel algorithms that solves fundamental problems in the standard workflow. Our simulation studies demonstrate that MuSiCal outperforms state-of-the-art algorithms for both signature discovery and assignment. By reanalyzing over 2,700 cancer genomes, we provide an improved reference catalog of signatures and their assignments, discover nine indel signatures absent in the current catalog, resolve long-standing issues with the ambiguous 'flat' signatures, and give insights into signatures with unknown etiologies. We expect MuSiCal and the improved catalog to be a step towards establishing best practices for mutational signature analysis.

1 Introduction

Based on the observation that different mutational processes leave unique patterns of somatic alteration on the DNA [1–5], mutational signature analysis has emerged as a powerful tool for unveiling these processes from DNA sequencing data, providing, for example, a glimpse into the mutational history of cancer development [6–12]. By analyzing a large collection of tumors profiled with whole-genome and whole-exome sequencing, a number of mutational signatures consisting of single-base substitutions (SBSs), indels (IDs), and structural variations (SVs) have been discovered [13–15]. Some of these signatures have been associated with specific exogenous or endogenous mutational processes and validated through experiments using CRISPR assays or mutagen-treated cell cultures [16–18], although the majority still have unknown etiologies. The application of mutational signature analysis has generated mechanistic insights into the mutational forces driving tumorigenesis [19, 20]. Early evidence shows that it may also be instrumental in informing prognosis and treatment decisions for cancer patients [21–27].

To maximize the impact of mutational signature analysis, rigorous mathematical understanding and robust algorithms are needed to increase its accuracy and interpretability. Despite the field's rapid development in the past decade, several methodological difficulties still remain [11, 28]. One such difficulty is in the signature assignment step, in which the set of active mutational signatures and their corresponding exposures are determined for either a dataset or a single sample. Erroneous signature assignment can produce misleading results, such as tobacco smoking-related signatures appearing in tumor types unlikely to be affected by tobacco smoking, or signatures associated with mismatch repair deficiency (MMRD) appearing in mismatch repair-proficient (MMRP) samples. Such misleading signature assignments are pervasive in the literature, as demonstrated using the results from two recent pan-cancer studies [29, 30] (Extended Data Fig. 1). Although the problem with well-annotated signatures is easier to recognize, it becomes more difficult for signatures with unknown etiologies. The relatively 'flat' signatures, such as SBS3, 5, 8, and 40, are particularly problematic since they are more difficult to separate from each other mathematically (Extended Data Fig. 2a) [14, 22, 31–34]. The same issue exists for signatures with similar spectra (Extended Data Fig. 2b). As

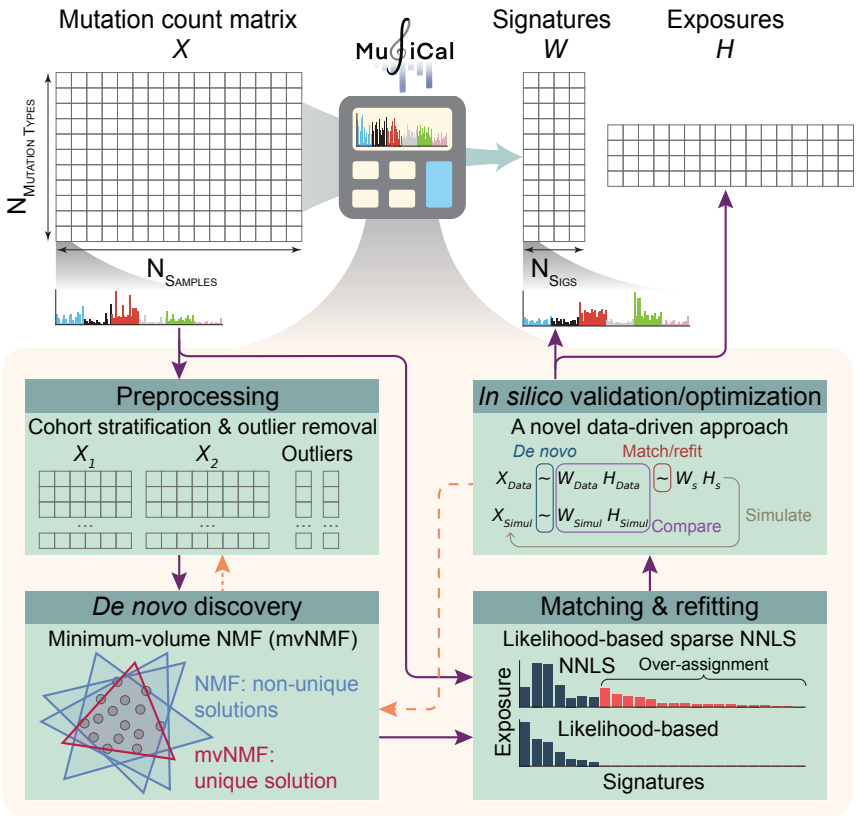


Fig. 1 Overview of MuSiCal. MuSiCal decomposes a mutation count matrix X into a signature matrix W and an exposure matrix H through four main modules – preprocessing, *de novo* discovery, matching and refitting, and *in silico* validation/optimization. The novel methods implemented in each module are highlighted.

a result, some of these signatures have to be combined to avoid uncertainties in their assignment [22, 33, 34], making it difficult to investigate their respective etiologies through association with clinical or genomic covariates. The importance of improving the signature assignment step cannot be overstated, as inaccurate signature assignment can have profound implications in therapeutic decision-making. For example, the observation of SBS3, which corresponds to homologous recombination deficiency (HRD) and suggests poly(ADP-ribose) polymerase (PARP) inhibitors as a treatment option [35, 36], in a significant proportion of multiple myelomas has been found to be illusive due to flawed signature assignment [22, 28, 37]. Other signatures have been proposed as biomarkers of therapy sensitivity, such as MMRD-associated signatures for immunotherapy [18, 38] and APOBEC-associated signatures for ATR inhibitors [39]. However, accurate signature-based patient stratification heavily depends on reliable signature assignment. In fact, difficulties in signature assignment may have undermined the performance of previous signature-based classifiers of patient subgroups and limited their clinical applications [29, 30].

Another difficulty is in the signature discovery step, in which mutational signatures are directly extracted from a cohort. Numerous tools have been developed, but they lead to highly variable results, even from the same dataset [14, 40]. It is therefore difficult to compare signatures discovered across studies, especially to determine whether a signature is novel or merely a variation of a known one due to algorithmic biases. This issue may underlie the large number of signatures that are overly similar or can be expressed as linear combinations of each other in the widely used catalog of known signatures from Catalog of Somatic Mutations in Cancer (COSMIC) [41] (Extended Data Fig. 2c). It may also confound the tissue specificity observed for some signatures [30].

In the conventional workflow, a mutation count matrix X is first collected by categorizing mutations from a cohort of samples into classes. X is then deconvolved into a signature matrix W and an exposure matrix H through *de novo* signature discovery. To obtain final signature assignments, a series of post-discovery steps are performed, including matching the *de novo* signatures to a known catalog (e.g., from COSMIC) and refitting. In 'refitting', a refined exposure matrix is determined by fitting X with matched signatures from the catalog as well as potential novel signatures that could not be matched. Most current approaches rely on non-negative matrix factorization (NMF) for *de novo* signature discovery [42] and non-negative least squares (NNLS) for matching and refitting.

Here, we identify methodological problems with the standard workflow, demonstrate that they contribute to the issues described above, and propose novel methods to solve these problems. We present Mutational Signature

Calculator (MuSiCal), a comprehensive framework that enables accurate signature assignment as well as robust and sensitive signature discovery (Fig. 1). MuSiCal leverages the minimum-volume regularization to solve the non-uniqueness problem of NMF-based *de novo* signature discovery [43–46], employs a novel likelihood-based sparse NNLS algorithm for matching and refitting to introduce proper sparsity constraints that are lacking in plain NNLS, and introduces a data-driven approach for systematic parameter optimization and *in silico* validation during post-discovery steps. In addition, to facilitate future studies of mutational signatures in cancer, we present an improved reference resource for both signatures and their assignments by reanalyzing over 2,700 genomes from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project [47]. Our new catalog of indel (ID) signatures is substantially more complete, containing 9 novel ID signatures absent in the current COSMIC catalog. The derived signature assignments also resolve issues with ambiguous flat signatures, providing insights into signatures with unknown etiologies.

2 Results

2.1 *De novo* mutational signature discovery with minimum-volume NMF

Although NMF is commonly used, it has an intrinsic problem of producing non-unique solutions [48–50]. This can be illustrated by a geometric interpretation of NMF (Fig. 2a): any cone that encloses the data points is a valid NMF solution, and there exist infinitely many such cones for a fixed dataset in general. Indeed, given the same input data, multiple NMF runs converge to different solutions, none of which recover the true signatures accurately (Fig. 2a). Relatively flat signatures (e.g., SigC in Fig. 2a) have especially large errors due to the lack of constraints. In Fig. 2b, we simulate profiles with real signatures SBS3, 8, 12, and 23 and try to recover them using NMF. Whereas SBS8, 12, and 23 are recovered accurately, the flatter SBS3 receives significant distortions. Specifically, the NMF solution for SBS3 has reduced weights in trinucleotide categories enriched in the other three signatures. As a result, it may be misidentified as a novel signature or, when decomposed using COSMIC signatures by NNLS, false signatures (e.g., SBS39) are introduced (Fig. 2b). Therefore, the non-uniqueness of NMF solutions can lead to algorithmic distortions that are difficult to distinguish from real biological signal, posing significant challenges to accurate signature discovery and assignment.

We propose a new approach based on minimum-volume NMF (mvNMF), originally developed for audio-source separation and hyperspectral-image unmixing [43–46]. Compared to NMF, mvNMF adds a regularization term that penalizes the volume of the cone formed by discovered signatures (i.e., area of the purple shaded triangle in Fig. 2a), promoting the unique solution with minimum volume among those that are equivalent based on NMF (see Methods for algorithmic details). In the examples above, mvNMF accurately recovers the true signatures without distortions (Fig. 2a, b).

For a more systematic comparison, we run NMF and mvNMF on synthetic datasets generated using tumor type-specific SBS signatures and Dirichlet-distributed exposures (Methods). Consistent with the examples in Fig. 2a, b, mvNMF improves the accuracy of discovered signatures compared to NMF, reducing the cosine error by 67-98% when averaged across different signatures within the same tumor type (Extended Data Fig. 3). The improvement is especially notable for relatively flat signatures (Fig. 2c). Even sparse signatures with nonzero weights only in a few SBS categories can pose substantial challenges for NMF because of the interference received from other similar signatures, whereas mvNMF does not suffer from these distortions (see, e.g., SBS7a in Extended Data Fig. 3d).

Theoretically, mvNMF is guaranteed to recover the true underlying signatures if the exposure matrix is sparse enough such that only a few signatures are active in each sample, whereas NMF requires much stronger assumptions [46, 50, 51]. Indeed, strong sparsity is observed for tumor type-specific exposure matrices produced by PCAWG [14], justifying the application of mvNMF to cancer data (Extended Data Fig. 4a-d). The increased accuracy of mvNMF holds for a wide range of exposure sparsity relevant for cancer data (Extended Data Fig. 4e). Even in the unrealistic case of non-sparse exposures, mvNMF ensures that the solutions are non-negative mixtures of true signatures, enabling the true underlying signatures to be uncovered through the downstream NNLS-based matching step (Extended Data Fig. 5). By contrast, NMF solutions have no such properties and can be arbitrarily distorted. These distortions may result in the discovery of false signatures in the matching step, leading to erroneous signature assignments (Extended Data Fig. 5, also evident from the example in Fig. 2b). Of note, another form of volume regularization has been recently applied to mutational signature discovery for germline mutations [52]. Unlike somatic mutations from tumor samples, the exposure matrix of germline mutations is in general non-sparse due to the lack of heterogeneity, especially when the mutations are analyzed across genomic bins as in [52]. Accordingly, volume regularization is enforced on the exposure matrix in [52]. By contrast, MuSiCal regularizes the volume of the signature matrix (Methods). Therefore, MuSiCal employs the volume regularization that is more suitable for tumor somatic mutations and represents the first application of mvNMF to cancer data.

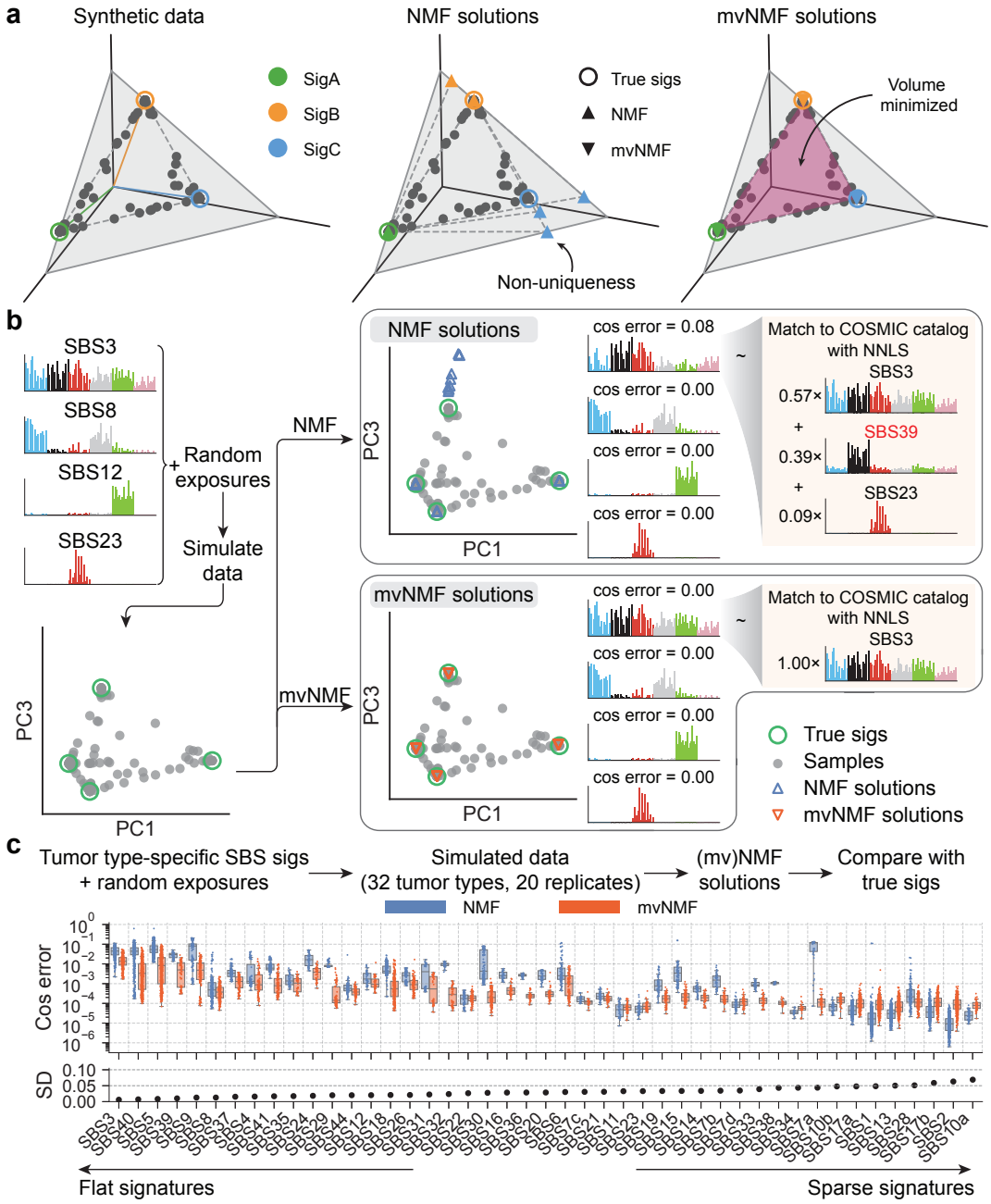


Fig. 2 Minimum-volume NMF (mvNMF) improves the accuracy of *de novo* signature discovery by inducing unique solutions. a. An example demonstrating the non-uniqueness of NMF solutions and uniqueness of mvNMF solutions. Synthetic samples are simulated from signatures with three mutation channels. NMF and mvNMF are then applied to recover the signatures with three different initializations (shared between NMF and mvNMF). Both signatures and samples are normalized and subsequently visualized on the plane $x + y + z = 1$. Geometrically, NMF looks for signatures such that the cone (represented by dashed triangles) formed by the signatures encloses all data points within its interior. Since infinitely many such cones exist, NMF solutions are not unique. By comparison, mvNMF further requires that the cone has minimum volume and thus promotes a unique solution. **b.** A similar example with real SBS signatures. Example solutions are plotted for NMF and mvNMF. The NMF solution for SBS3 receives a relatively large cosine error, and when decomposed using COSMIC signatures by NNLS (with a relative exposure cutoff of 0.05), it is identified as a composite signature involving the false SBS39. By comparison, mvNMF is able to recover SBS3 accurately. **c.** Cosine errors of NMF and mvNMF solutions in tumor type-specific simulations with real SBS signatures. mvNMF improves the accuracy of *de novo* signature discovery for most SBS signatures, especially for relatively flat ones characterized by small standard deviations (SD, shown below the box plot). Although NMF outperforms mvNMF for some sparse signatures, the cosine errors for both algorithms are smaller than 0.001 in those cases and thus negligible.

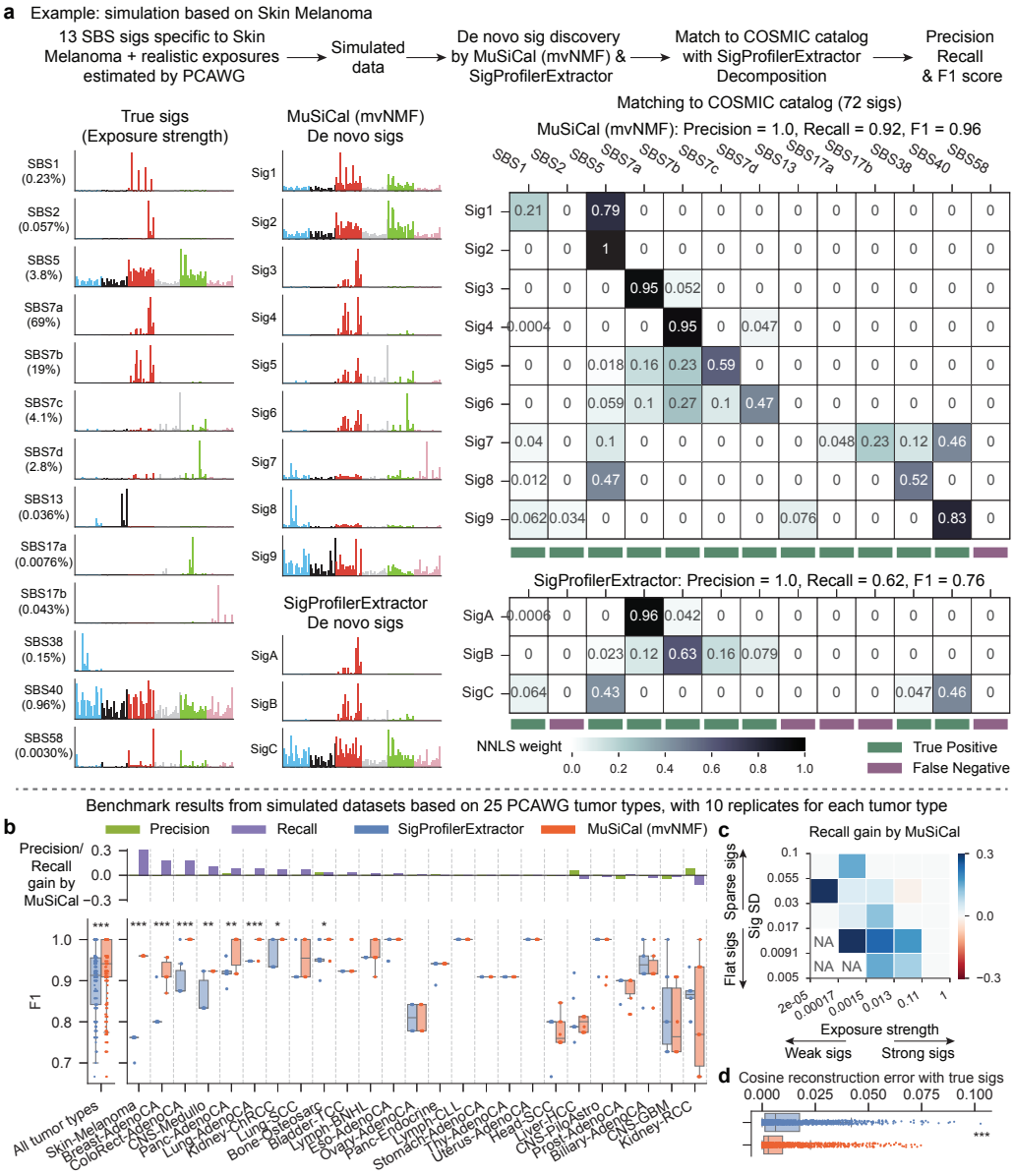


Fig. 3 MuSiCal outperforms the state-of-the-art algorithm SigProfilerExtractor for *de novo* signature discovery. a. An example based on Skin-Melanoma demonstrating the metric for evaluating the quality of *de novo* signature discovery. A synthetic dataset is simulated from 13 SBS signatures specific to Skin-Melanoma (left) and the exposure matrix produced by the PCAWG consortium. MuSiCal and SigProfilerExtractor are applied to derive *de novo* signatures (middle), which are subsequently decomposed as non-negative mixtures of COSMIC signatures (right). Precision, recall, and F1 score are then calculated by comparing matched COSMIC signatures with the 13 true signatures. **b.** Comparison of MuSiCal and SigProfilerExtractor for *de novo* signature discovery from synthetic datasets based on 25 PCAWG tumor types with at least 20 samples. 10 independent simulations are performed for each tumor type. Box plot shows F1 score for each individual dataset. Bar plot shows the mean precision and recall gain by MuSiCal, as compared to SigProfilerExtractor. Tumor types are sorted according to the mean F1 gain by MuSiCal. **c.** Recall gain by MuSiCal for signatures of different exposure strengths and sparsity levels, showing that MuSiCal is especially sensitive for discovering weak signatures. Exposure strengths are not directly comparable between flat and sparse signatures. For example, 100 mutations may be considered a strong exposure for SBS2, which has prominent weights only at 2 of the 96 SBS categories, while the same number of mutations can only be considered a weak exposure for the flatter SBS3. Therefore, different signatures are grouped by exposure strength as well as sparsity, as quantified by the standard deviation (SD) of their corresponding spectra. Recall gain by MuSiCal, as compared to SigProfilerExtractor, is then calculated for signatures within each group separately. **d.** Cosine reconstruction errors for MuSiCal- and SigProfilerExtractor-derived *de novo* signatures. Each *de novo* signature is decomposed into a non-negative mixture of the true underlying signatures with NNLS. Cosine distance is then calculated between the reconstructed and the original *de novo* signature. *: $p < 0.05$. **: $p < 0.005$. ***: $p < 0.0005$. p-values are calculated with paired t-tests.

2.2 MuSiCal outperforms the state-of-the-art algorithm for *de novo* signature discovery

We benchmarked the performance of MuSiCal for *de novo* signature discovery against the state-of-the-art algorithm SigProfilerExtractor [53]. To recapitulate the complexities of real-life data (such as different signature strengths,

correlated exposures, and small sample numbers and/or mutation counts), synthetic datasets are simulated from tumor type-specific SBS signatures and real exposure matrices produced by PCAWG [14] (Methods).

To infer the number of signatures r from the dataset itself, MuSiCal employs a novel parameter-free method (Methods, Extended Data Fig. 6). Previous approaches usually perform signature discovery for different r values and choose the one with small reconstruction errors and stability [7, 30, 53]. However, the choice is either manual or heavily dependent on predefined thresholds. We propose to cluster the solutions from multiple mvNMF runs with the same r via hierarchical clustering and then determine the optimal number of clusters k using the well-established gap statistic [54, 55]. We observe that when r is misspecified, mvNMF tends to produce unstable or redundant signatures, leading to $k \neq r$ (see, for example, Extended Data Fig. 6c, g, h). We thus select the largest r among those with $k = r$ as the number of signatures.

To evaluate the quality of *de novo* signature discovery, we use the accuracy of the signature assignment from the downstream matching step. We illustrate this metric with an example simulated from the 13 SBS signatures and real exposures produced by PCAWG for Skin-Melanoma (Fig. 3a). The exposure strengths of these true signatures (defined as the percentage of total mutation count contributed by a particular signature) vary greatly in the PCAWG result, ranging from 69% for SBS7a, associated with ultraviolet (UV) light exposure, to a mere 0.003% for SBS58, annotated as a potential sequencing artifact according to COSMIC. It is impractical to expect any algorithm to directly output precisely the 13 *de novo* signatures matching the underlying truth. Indeed, SigProfilerExtractor discovers only 3 *de novo* signatures, while MuSiCal discovers only 9 (Fig. 3a). Further inspection shows that most *de novo* signatures are mixtures of a subset of the 13 true signatures, potentially due to their strong co-occurring patterns (e.g., among the four UV-related signatures SBS7a-d), making direct comparison between the *de novo* and the true signatures difficult. We therefore perform a downstream matching step where each *de novo* signature is decomposed into a non-negative mixture of potentially multiple COSMIC signatures. We then define those COSMIC signatures to which at least one *de novo* signature matches as predicted positives and the rest of the COSMIC signatures as predicted negatives. Precision, recall, and their harmonic mean (the F1 score) can subsequently be calculated. For the Skin-Melanoma example, the 9 *de novo* signatures from MuSiCal are matched to 12 of the 13 true signatures, except the weakest signature SBS58. We thus consider 12 of the 13 true signatures to be successfully discovered, leading to a recall of $12/13 = 0.92$ (Fig. 3a). For SigProfilerExtractor, 4 fewer signatures are discovered, leading to a recall of $8/13 = 0.62$. For both algorithms, precision is 1 since no *de novo* signatures are matched to signatures other than the 13 true ones. Therefore, in this example, MuSiCal achieves an F1 score of 0.96, outperforming SigProfilerExtractor (F1 score = 0.76).

When we apply the same metric to all 25 tumor types (each with 10 independent simulations), MuSiCal obtains higher mean F1 scores in 14 tumor types (8 are statistically significant) compared to SigProfilerExtractor, driven by improved recall and without loss of precision (Fig. 3b). Notably, in several tumor types where SigProfilerExtractor produces suboptimal results, such as Skin-Melanoma, Breast-AdenoCA, and ColoRect-AdenoCA, MuSiCal achieves pronounced improvement, increasing the mean F1 score by 0.20, 0.11, and 0.10, respectively (the mean recall increases by 0.32, 0.18, and 0.18, respectively, while precision is similar). In the remaining 11 tumor types, MuSiCal obtains similar mean F1 scores (4 equal, 7 lower, and none statistically significant). When averaged across all tumor types, MuSiCal outperforms SigProfilerExtractor in mean F1 score (0.93 vs. 0.90, paired t-test $p < 10^{-6}$) and mean recall (0.88 vs. 0.84, paired t-test $p < 10^{-6}$), with similar precision (0.99 vs. 0.99). The seemingly small gains in F1 score and recall do not fully reflect the improvement obtained by MuSiCal, as these statistics are dominated by strong signatures that are relatively easy to discover. We thus further stratified the signatures based on their exposure strengths and investigated them separately. As shown in Fig. 3c, MuSiCal outperforms SigProfilerExtractor especially for discovering weak signatures, increasing recall by up to 0.3. This marked increase in sensitivity for weak signatures is also evident from the Skin-Melanoma example in Fig. 3a: SigProfilerExtractor fails to discover the 5 weakest signatures, whereas MuSiCal rescues 4 of the 5 undetected signatures and misses only the weakest signature SBS58.

These results demonstrate that MuSiCal outperforms the state-of-the-art algorithm for *de novo* signature discovery in simulations emulating real-life datasets. With its higher sensitivity for weak signatures, MuSiCal could facilitate the discovery of novel signatures missed by previous methods. The improved performance of MuSiCal relies on the reduced algorithmic bias powered by mvNMF, as indicated by reduced cosine reconstruction errors of the *de novo* signatures when decomposed using the true signatures (Fig. 3d). Of note, the same matching algorithm was used for both MuSiCal- and SigProfilerExtractor-derived *de novo* signatures to ensure a fair comparison (see Methods and Extended Data Fig. 7 for more details).

2.3 Refitting with likelihood-based sparse non-negative least squares (NNLS)

Another key improvement in MuSiCal is in the refitting step, in which an observed mutational spectrum is decomposed into a combination of existing signatures with non-negative coefficients. Without additional constraints, NNLS

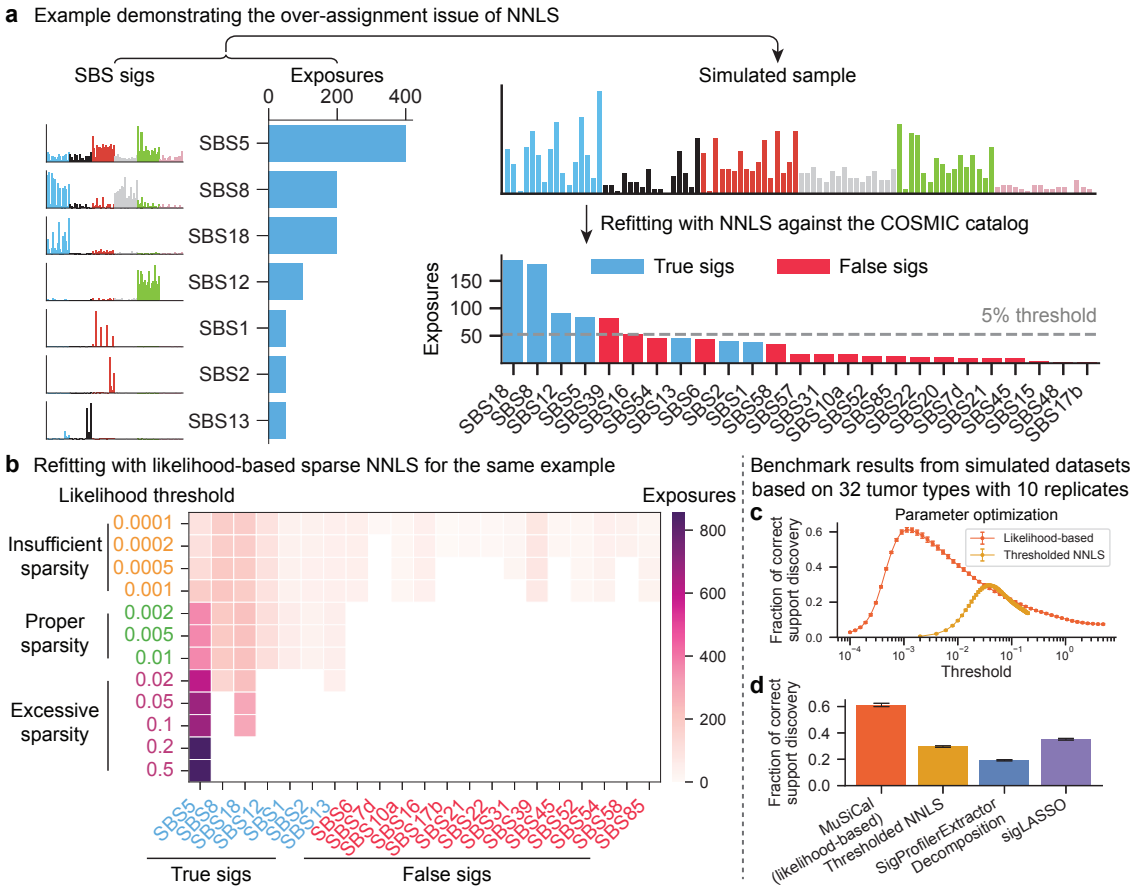


Fig. 4 MuSiCal solves the over-assignment problem of refitting using multinomial likelihood-based sparse NNLS. a. An example demonstrating the over-assignment problem of NNLS. A synthetic sample is simulated from 7 SBS signatures and a fixed exposure vector, and then fed into NNLS for refitting against the COSMIC catalog. As a result, 18 additional false positive signatures receive nonzero exposures. Even with a sparsity-inducing threshold, NNLS fails to discover the set of active signatures correctly. **b.** Solutions of likelihood-based sparse NNLS for the same example with different likelihood thresholds. With a properly chosen threshold, likelihood-based sparse NNLS identifies the set of active signatures correctly. The solutions also show a desired property of continuity with varying thresholds: when the threshold is too small, all 7 true signatures are discovered, as well as additional false positives; when the threshold is too large, only the strongest true signatures are discovered. **c.** Threshold tuning for likelihood-based sparse NNLS and thresholded NNLS on simulated datasets. The optimal threshold is chosen to maximize the correct support discovery rate. Error bar indicates standard deviation from 10 independent simulations. **d.** Performance of likelihood-based sparse NNLS and three other algorithms on simulated datasets. Results from MuSiCal (i.e., likelihood-based sparse NNLS) and thresholded NNLS are shown with optimal thresholds. Error bar indicates standard deviation from 10 independent simulations.

attempts to utilize all signatures in the catalog to minimize the fitting error, which often results in over-assignment (that is, a signature that should not be present gets a weight). We illustrate this problem with a simple example simulated from 7 SBS signatures (Fig. 4a). When refit against the COSMIC catalog with NNLS, 18 additional signatures receive nonzero exposures, some of which have even higher exposures than the true signatures. A common method to introduce sparsity in refitting is thresholded NNLS, where signatures that contribute less than a specified percentage (e.g., 5%) to the total mutation burden are set to have zero exposures [56]. The problem with a fixed threshold is that different mutational signatures require different numbers of mutations to provide confidence in their presence. Some signatures may be statistically significant even with a small number of mutations, whereas others may not be significant even with many mutations. For example, 20 C>T mutations at CpGs may be sufficient to claim the presence of SBS1, which has prominent weights only at 4 of the 96 SBS categories, whereas the same number of mutations can provide only limited evidence for the flatter SBS3. Indeed, in the simulated example above, there is no fixed threshold that recovers the 7 true signatures without adding false positives (Fig. 4a).

We propose a sparse NNLS algorithm based on multinomial likelihoods (Methods, Extended Data Fig. 8a). We first perform NNLS to obtain an initial exposure estimate and then refine the exposure vector in a bidirectional stepwise manner. During the backward step, a currently active signature is excluded if removing it does not significantly reduce the multinomial likelihood of the observed mutation count vector. Similarly, during the forward step, a currently inactive signature is included if adding it significantly increases the likelihood. We repeat these two steps until no

removal (or addition) of a signature results in a decrease (or increase) of the multinomial likelihood below (or above) a threshold. The use of multinomial likelihoods provides a rigorous way to assess the presence of a signature, regardless of its flatness.

In the example above, when the likelihood thresholds are varied, solutions produced by likelihood-based sparse NNLS show a desired property of continuity: with relatively small thresholds, all 7 true signatures are discovered, as well as additional false positives; with relatively large thresholds, only the strongest true signatures are discovered (Fig. 4b). Importantly, there exists a range of thresholds where all 7 true signatures are discovered without any false positives (Fig. 4b). Therefore, with a proper threshold (see next section on how to choose this threshold), likelihood-based sparse NNLS is able to identify the set of active signatures correctly in this example, alleviating the over-assignment problem. Of note, the widely used cosine similarity is less powerful in separating similar signatures compared to multinomial likelihood (Extended Data Fig. 8b) and fails to discover the set of active signatures correctly at any threshold (Extended Data Fig. 8c).

For systematic evaluation, we simulate synthetic datasets with tumor type-specific SBS signatures (Methods). Refitting is performed against the entire COSMIC catalog with likelihood-based sparse NNLS, thresholded NNLS, as well as two state-of-the-art algorithms – sigLASSO [57] and the Decomposition module of SigProfilerExtractor [53]. Results are evaluated using the metric of correct support discovery rate, defined as the proportion of samples that have the correct set of active signatures identified. For the purpose of this simulation study, the thresholds for likelihood-based sparse NNLS and thresholded NNLS are tuned such that the best correct support discovery rate is achieved (Fig. 4c), although in practice, MuSiCal optimizes the likelihood threshold through the *in silico* validation/optimization module without knowing the underlying truth (see next section). With the optimal threshold, likelihood-based sparse NNLS outperforms all three other methods, increasing the correct support discovery rate by 74-218% (61% vs. 35-19%) (Fig. 4d).

Finally, MuSiCal uses the same likelihood-based sparse NNLS for the matching step, where *de novo* signatures are decomposed as non-negative mixtures of known signatures in the catalog (Methods, Extended Data Fig. 8d).

2.4 Data-driven parameter optimization and *in silico* validation

To improve the interpretability of final results, *de novo* signature discovery is usually followed by post-processing steps, including matching and refitting. Currently, the post-processing steps are highly variable across different tools, leading to vastly different final signature assignments even from the same *de novo* discovery results. In particular, two crucial problems have hindered the accuracy of signature assignment. First, various parameters involved in post-processing are arbitrarily chosen, manually tuned, or selected based on simplistic simulations that fail to capture the complexity of real data. Second, there is no systematic way of testing the consistency between final signature assignment and data, or quantifying the lack thereof. The simple metric of reconstruction error is inadequate since it does not inform overfitting or signature misassignment.

To tackle these problems, we propose a novel data-driven approach for *in silico* validation (Methods, Fig. 5a). Specifically, we simulate a synthetic dataset X_{simul} from the final signature assignments $(W_{\text{s}}, H_{\text{s}})$ (s stands for sparse), which are obtained through post-processing steps from the *de novo* discovery results $(W_{\text{data}}, H_{\text{data}})$. The simulated data X_{simul} then undergoes *de novo* signature discovery with exactly the same settings (including the selected number of signatures r) as used for the original data X_{data} to derive $(W_{\text{simul}}, H_{\text{simul}})$. We reason that if $(W_{\text{s}}, H_{\text{s}})$ faithfully describes X_{data}, the *de novo* discovery results from simulations and data should be consistent with each other. We thus compare $(W_{\text{simul}}, H_{\text{simul}})$ with $(W_{\text{data}}, H_{\text{data}})$ and use the resulting distances to quantify the consistency between the signature assignments $(W_{\text{s}}, H_{\text{s}})$ and the original data X_{data}. These distances provide a systematic way to validate the final signature assignments *in silico*. Of note, the key innovation is to compare the *de novo* discovery results instead of X_{simul} and X_{data}, which merely reflect reconstruction errors. Doing the comparison immediately after *de novo* discovery also avoids any influence from potential biases in the post-processing steps, allowing us to assess the quality of post-processing and thus optimize the parameters therein (see below).

To illustrate that the validation scheme described above is able to assess the quality of signature assignment and reveal potential issues, we study the PCAWG glioblastoma dataset as an example and compare the final signature assignments from MuSiCal and the PCAWG consortium [14] (Fig. 5b-f). MuSiCal discovers 4 *de novo* signatures in this dataset and with matching and refitting, assigns 8 COSMIC signatures (SBS1, 2, 5, 8, 12, 13, 30, and 31), whereas PCAWG assigns 4 COSMIC signatures (SBS1, 5, 30, and 40) (Fig. 5b). Following the validation scheme, synthetic datasets are simulated from both the MuSiCal and PCAWG assignments and undergo *de novo* signature discovery again. We find that the signature assignments from MuSiCal show higher consistency between simulation and data than those from PCAWG (Fig. 5c, d), suggesting improved accuracy of MuSiCal assignments for describing the original data. To investigate potential issues with the PCAWG assignment, *de novo* signatures from simulations and data are decomposed using the corresponding assigned signatures, and the resultant NNLS weights are compared (Fig.

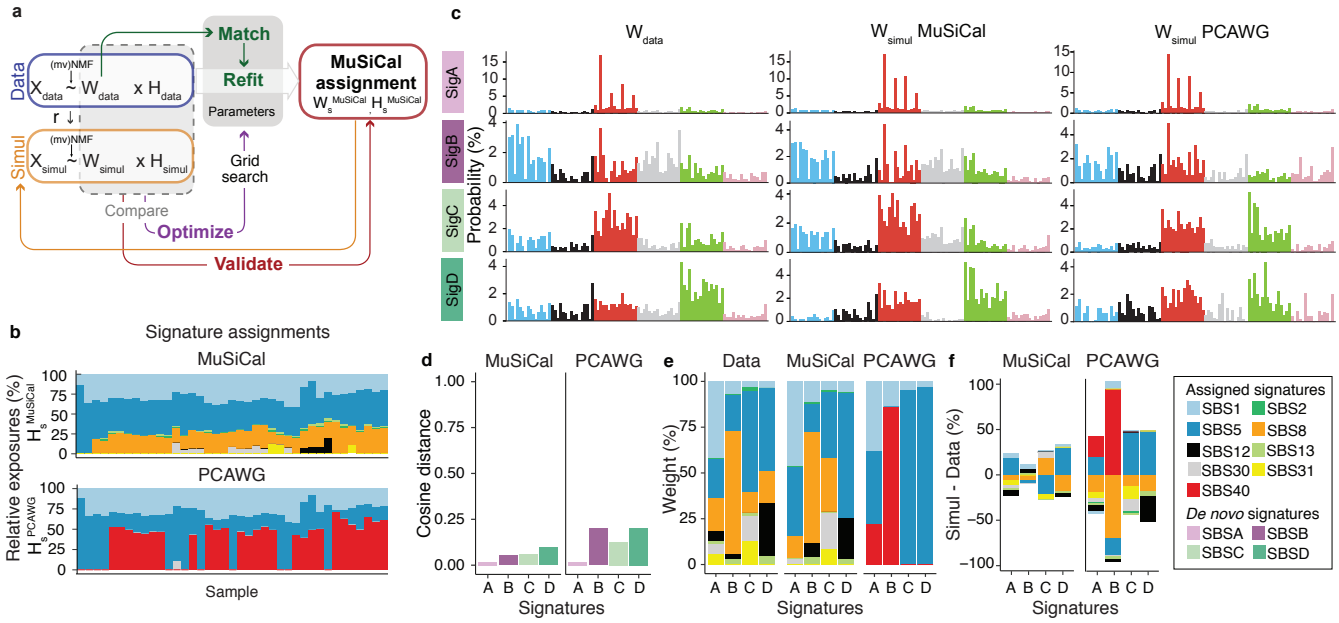


Fig. 5 ***In silico* validation quantifies the consistency between signature assignments and data. a.** Diagram of the data-driven approach for parameter optimization and *in silico* validation of post-processing steps after *de novo* signature discovery. See main text for details. **b.** Signature assignments by MuSiCal (top, H_s^{MuSiCal}) and the PCAWG consortium (bottom, H_s^{PCAWG}) [14] for the PCAWG glioblastoma dataset, showing discrepancies in both assigned signatures and the corresponding exposures. For simplicity, samples with MMRD or hypermutation due to temozolomide treatment are removed. **c.** Four *de novo* signatures discovered from the original data (left, W_{data}) are compared to those discovered from simulated data based on MuSiCal (middle, $W_{\mathrm{simul}}^{\mathrm{MuSiCal}}$) and PCAWG (right, $W_{\mathrm{simul}}^{\mathrm{PCAWG}}$) assignments. **d.** Cosine distances between W_{data} and $W_{\mathrm{simul}}^{\mathrm{MuSiCal}}$ (left) are compared to those between W_{data} and $W_{\mathrm{simul}}^{\mathrm{PCAWG}}$ (right), showing that MuSiCal assignments achieve better consistency between simulation and data in terms of *de novo* signatures. **e.** NNLS weights of *de novo* signatures for both simulations and data. *De novo* signatures from data (W_{data}) and MuSiCal-derived simulations ($W_{\mathrm{simul}}^{\mathrm{MuSiCal}}$) are matched to the MuSiCal-assigned signatures, i.e., SBS1, 2, 5, 8, 12, 13, 30, and 31. *De novo* signatures from PCAWG-derived simulations ($W_{\mathrm{simul}}^{\mathrm{PCAWG}}$) are matched to the PCAWG-assigned signatures, i.e., SBS1, 5, 30, and 40. Standard NNLS is used and the obtained weights are plotted. **f.** Differences between the NNLS weights in (e) for the original data and those for MuSiCal- and PCAWG-derived simulations. Positive and negative differences are plotted separately in opposite directions in a cumulative manner. The results are indicative of over-assignment of SBS40 and under-assignment of SBS8 by PCAWG as well as much improved signature assignment by MuSiCal.

5e, f). As expected, a large discrepancy is observed for the decomposition results between data and PCAWG-derived simulations. In particular, the *de novo* signatures from PCAWG-derived simulations contain strong components of SBS40, which lacks any evidence from data, but find no matches to SBS8, which has strong evidence from data. These results are indicative of over-assignment of SBS40 and under-assignment of SBS8 by PCAWG for the glioblastoma dataset. In fact, as we will demonstrate later, similar misassignment problems, especially of SBS40, are prevalent in the results produced by the PCAWG consortium [14].

Finally, we use the same data-driven approach to optimize post-processing parameters. In MuSiCal, these parameters are the two likelihood thresholds for likelihood-based sparse NNLS in matching and refitting, respectively. The optimization is achieved through a grid search. Specifically, among the pairs of thresholds that minimize the distance between W_{data} and W_{simul}, the pair with the sparsest signature assignment is selected (Methods, Extended Data Fig. 9).

2.5 Deriving a reference of signatures and signature assignments in cancer by reanalyzing PCAWG data

Empowered by the improved accuracy and sensitivity of MuSiCal, we reanalyzed over 2,700 cancer genomes from PCAWG for both SBS and indel (ID) signatures to derive a refined reference of signatures and their assignments (Methods). Tumor type-specific *de novo* signature discovery is performed after stratifying heterogeneous datasets and removing outliers with the preprocessing module (Extended Data Fig. 10). The discovered signatures are subsequently matched to the COSMIC catalog, and potential novel signatures are identified when a reasonable match cannot be found. Refitting then follows to estimate per-sample exposures. The likelihood thresholds in both matching and refitting are optimized via a grid search through data-driven simulations. The consistency between final signature assignments and data are assessed with the *in silico* validation module.

ID signatures can provide unique information on both endogenous and exogenous mutagenic processes, but they have been under-studied until recently, in part due to the relatively low mutational burden of IDs compared to SBSs

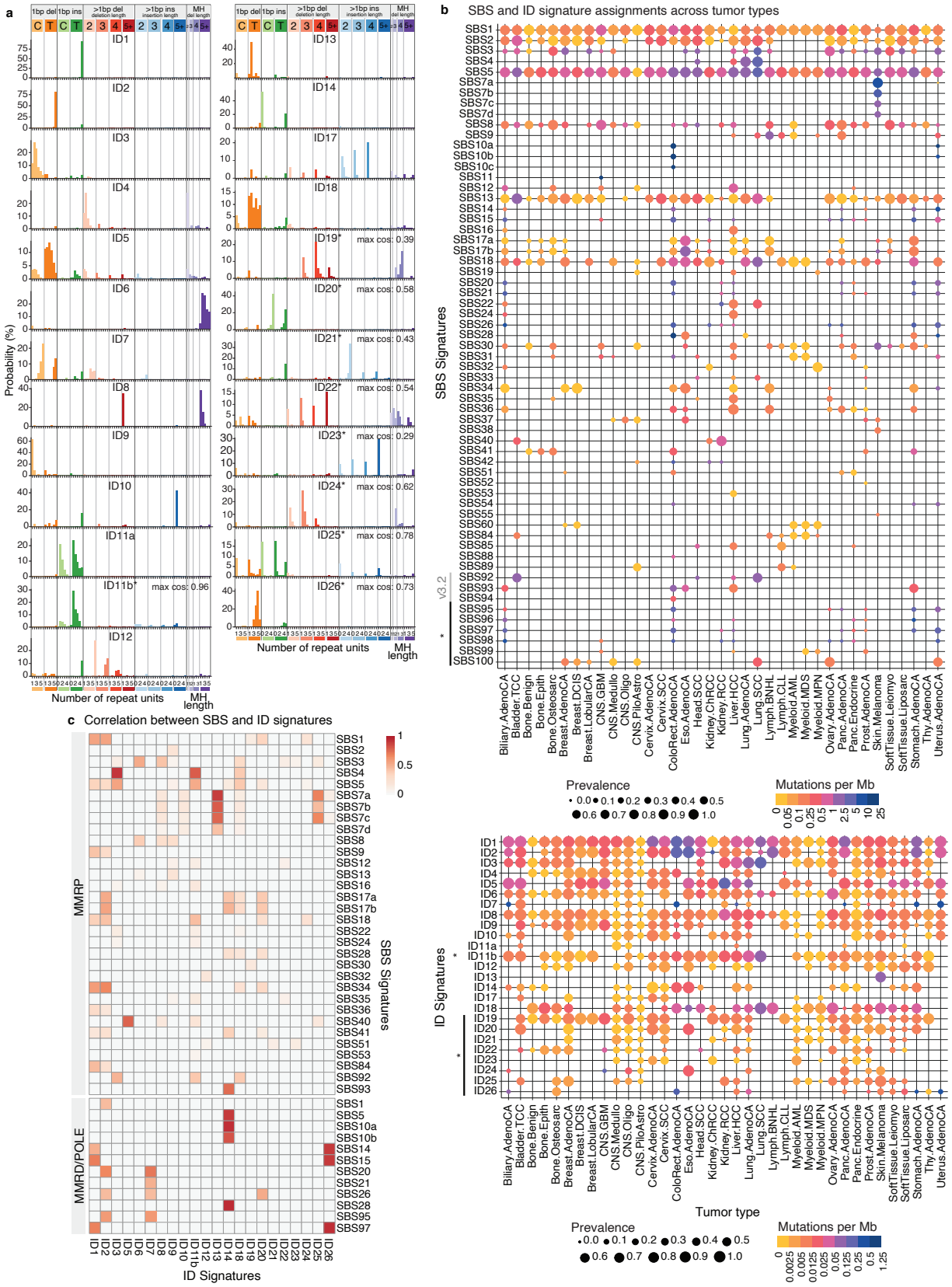


Fig. 6 (Continued on the following page.)

[14, 16, 18, 58–60]. Our reanalysis results in a substantial revision and expansion of the COSMIC catalog of ID signatures. With ID signatures, indels are classified into 83 different categories characterized by indel size, affected

Fig. 6 Reanalysis of PCAWG data with MuSiCal uncovers novel signatures and SBS-ID associations. a. The catalog of ID signatures obtained from our reanalysis of PCAWG data with MuSiCal. * indicates newly added signatures not present in the current COSMIC catalog. Other signatures also have modified spectra. Each of the 9 new signatures is decomposed using COSMIC signatures by NNLS. The resulting cosine similarities between the original and the reconstructed signatures are annotated next to each spectrum. **b.** MuSiCal-derived signature assignments in PCAWG samples for SBS (top) and ID (bottom) signatures. Marker size represents the prevalence of a signature, i.e., the proportion of samples with nonzero exposures of the corresponding signature within a tumor type. Color indicates the median number of mutations per Mb contributed from the corresponding signature among samples with nonzero exposures. **c.** Heatmap of Pearson correlation coefficients between the per-sample exposures of SBS and ID signatures. The correlations are calculated for MMRP samples and samples with MMRD and/or POLE-exo mutations separately. Only the samples with at least 100 indels are used and only the statistically significant associations are shown (adjusted $p < 0.05$). The correlations are also examined separately in each tumor type where the corresponding ID and SBS signatures are both active, and are excluded if they are significant only in a small fraction ($< 20\%$) of these tumor types.

nucleotides, and the presence of repetitive and/or microhomology (MH) regions [14]. The current COSMIC catalog contains 18 ID signatures labeled ID1-18. 16 of these signatures are discovered with modified spectra in our reanalysis (ID1-14, 17, and 18) (Fig. 6a), whereas the other two (ID15 and 16) are not (Extended Data Fig. 11a). ID15 and 16 were found by the PCAWG consortium in three whole-exome sequenced samples from The Cancer Genome Atlas (TCGA) (Extended Data Fig. 11b). However, further inspection shows that ID15 and 16 likely reflect artifactual indels, as these three samples have exceptionally large indel counts (> 1000) with only small SBS counts (< 250) (Extended Data Fig. 11c). Moreover, their indel variant allele frequencies (VAFs) are much lower than their SBS VAFs, unlike in other TCGA samples (Extended Data Fig. 11d).

In addition, our reanalysis discovers 9 new ID signatures that are not in the COSMIC catalog (Fig. 6a). Specifically, 8 of the 9 have a cosine similarity < 0.8 when decomposed using COSMIC signatures, and we named them ID19-26. The remaining signature is similar to COSMIC ID11 (cosine similarity $= 0.94$) and we refer to it as ID11b. Compared to COSMIC ID11, which is characterized by 1-bp insertions of C and T at their respective mononucleotide repeats, ID11b is dominated by T insertions and thus included as a new signature. Together, we constructed an updated catalog of ID signatures by replacing COSMIC ID1-14, 17, and 18 with the corresponding signatures from our reanalysis (ID11 renamed to ID11a to distinguish from ID11b), removing the likely artifactual COSMIC ID15 and 16, and adding the new ID11b and ID19-26 (Fig. 6a, Supplementary Table 1). We note that it is necessary to replace the existing COSMIC spectra with MuSiCal-derived results because several COSMIC ID signatures are resolved into multiple signatures in our updated catalog. For example, COSMIC ID4 is separated into MuSiCal ID4, 19, and 24 (Extended Data Fig. 11e). Interestingly, our updated MuSiCal version of ID4 is very similar to the experimentally derived signature of topoisomerase 1 (TOP1)-mediated deletions in RNase-H2-null cells, predominantly consisting of 2-bp deletions at short tandem repeats (STRs) with 2 or 3 repeat units, as well as 2-bp deletions with 1-bp MH [58] (Extended Data Fig. 11f). By comparison, the COSMIC version of ID4 contains significant proportions of longer (3- and 4-bp) deletions that are not observed in these experiments, suggesting that COSMIC ID4 is indeed a potential mixture of multiple signatures (Extended Data Fig. 11f).

For SBSs, most signatures discovered from our reanalysis are successfully matched to the current COSMIC catalog (SBS1-60, and SBS84-94). The 6 new SBS signatures with relatively poor matches are mainly (4 of the 6) discovered in MMRD tumors, likely representing nonlinear interactions between DNA damage and repair [61] (Extended Data Fig. 12). We thus constructed an updated catalog of SBS signatures by adding the 6 new ones named SBS95-100, while keeping the COSMIC SBS signatures unchanged (Supplementary Table 2). A global update (as for ID signatures) was not performed since no substantial difference was observed between COSMIC and MuSiCal spectra. One exception is SBS40, a relatively flat signature annotated as having unknown etiology and as being correlated with patients' age in some tumor types. The MuSiCal-derived SBS40 has higher proportions of T>N mutations (57%) compared to the COSMIC SBS40 (41%) (Extended Data Fig. 13). Therefore, we replaced COSMIC SBS40 with this modified spectrum in the updated SBS catalog.

With the updated catalogs of ID and SBS signatures, tumor type-specific signature assignments display marked differences compared to results obtained by the PCAWG consortium [14] (Fig. 6b, Extended Data Fig. 14, Supplementary Table 3 and 4). We elaborate on these differences in later sections.

2.6 Novel associations between SBS and ID signatures inform unknown etiologies

Of the 25 ID signatures we identified with MuSiCal, 16 (other than the 9 with annotations from COSMIC) have unknown etiologies. To infer their etiologies, we investigate the association between ID and SBS signatures, utilizing the large number of SBS signatures with annotated etiologies (41 from COSMIC). As many mutagenic processes produce IDs and SBSs simultaneously in experimental studies [16, 18], co-occurrence of ID and SBS signatures can indicate a common origin. In Fig. 6c, we show the statistically significant correlations between per-sample exposures of ID and SBS signatures obtained by MuSiCal for PCAWG data. Correlations are calculated separately for MMRP

samples and samples with MMRD and/or polymerase epsilon exonuclease (POLE-exo) domain mutations, since hypermutations in the latter group can confound the correlation analysis of MMRP samples.

As expected, strong correlations are observed between ID and SBS signatures with known and related etiologies. ID3 correlates with SBS4, both of which are associated with tobacco smoking (Fig. 6c). The characteristic C deletions at 2- or 3-bp mononucleotide C repeats (i.e., CC or CCC) in ID3 potentially result from repairs of G adducts formed by tobacco carcinogens (Fig. 6a) [16]. ID6 correlates with SBS3, producing ≥5-bp deletions with overlapping microhomology (MH), consistent with elevated activities of MH-mediated end joining (MMEJ) in HRD tumors (Fig. 6a, c). ID13, dominated by T deletions at TT dinucleotides, correlates with SBS7 and is associated with UV light exposure (Fig. 6a, c). Several ID signatures, including ID1, 2, and 7, are correlated with MMRD-related SBS14, 15, 20, 21, and 26 (Fig. 6c). ID7 primarily comprises 1- and 2-bp deletions at long repetitive sequences and is known to contribute to a large number of indels in MMRD tumors (Fig. 6a) [14]. ID1 and 2, characterized by 1-bp T insertions and deletions at long homopolymers of T, respectively, reflect higher levels of replication slippage mutagenesis in MMRD tumors (Fig. 6a) [14, 18].

Novel correlations are also observed for both newly added ID signatures in our updated catalog and existing signatures with revised spectra. ID11b correlates with both SBS4 and ID3 (Fig. 6c, Extended Data Fig. 15a, b), suggesting a potential origin of tobacco smoking-related mutagenesis. ID11b is composed predominantly of 1-bp T insertions following a T, with additional contributions from T insertions following TT or TTT (Fig. 6a). Similar T insertions, as well as the C deletions at CC or CCC observed in ID3, are seen in cells exposed to tobacco carcinogens dibenz[a,h]anthracene (DBA) and its diol-epoxide metabolite dibenz[a,h]anthracene diol-epoxide (DBADE) (Extended Data Fig. 15c) [16]. By comparison, another two components of tobacco smoke – benzo[a]pyrene (BaP) and its diol-epoxide metabolite benzo[a]pyrene-7,8-dihydrodiol-9,10-epoxide (BPDE) – result in only ID3 and not ID11b (Extended Data Fig. 15c) [16]. Interestingly, all four chemicals produce SBS signatures similar to SBS4 (Extended Data Fig. 15c) [16]. In smoking-related (i.e., SBS4-positive) lung, liver, and head-and-neck cancers, ID11b is indeed required to explain the observed abundance of 1-bp T insertions following a T (Extended Data Fig. 15d). Together, these results suggest that ID11b represents another component of tobacco smoking-related mutagenesis and is likely specific to DNA damages produced by DBA and DBADE.

ID14, dominated by 1-bp C insertions at non-repetitive regions and 1-bp T insertions at long (≥ 5-bp) mononucleotide T repeats, correlates with SBS10 and 28, indicating that ID14 is associated with POLE-exo mutations (Fig. 6a, c). ID26, characterized by 1-bp T deletions at 4- and 5-bp mononucleotide T repeats and potentially resolved from COSMIC ID7, correlates with SBS14 and 15, suggesting an MMRD-related origin (Fig. 6a, c). Finally, ID25, consisting of mostly 1-bp indels at non-repetitive regions and T indels at long mononucleotide T repeats, correlates with SBS7 and contributes to relatively large numbers of mutations in skin melanomas, suggesting that DNA damage by UV light exposure or related repair processes may underlie ID25 (Fig. 6a, b, c).

Some of the SBS-ID correlations mentioned above are also observed with exposure matrices produced by the PCAWG consortium. However, the results from MuSiCal show stronger correlations overall and reveal novel ones that are not detected by PCAWG (Extended Data Fig. 16).

2.7 Improved signature assignment resolves ambiguities with flat signatures

To demonstrate the improved signature assignments by MuSiCal (Fig. 6b, Extended Data Fig. 14), we focus on SBS40. SBS40 is relatively flat, similar to SBS3 and 5 (cosine similarities = 0.88 and 0.83, respectively), and overlaps with many other signatures. Distinguishing SBS40 from other signatures and quantifying its contribution in different samples are thus mathematically difficult. PCAWG assigns SBS40 to nearly every tumor type (Extended Data Fig. 14) [14]. However, we believe this result is artifactual and caused by the over-assignment problem described earlier. Some studies have noticed this problem and have manually combined SBS40 with other flat signatures [33, 34]. Indeed, our *in silico* validation using data-driven simulations (described in Fig. 5a) confirms this over-assignment issue (Fig. 7a), as *de novo* signatures extracted from simulated datasets based on PCAWG assignments (i.e., $W_{\text{simul}}^{\text{PCAWG}}$) contain an excess amount of SBS40 (red bars in Fig. 7a) in most tumor types when compared to *de novo* signatures extracted from the original data (i.e., W_{data}). By comparison, MuSiCal assigns SBS40 only to kidney and bladder cancers (Fig. 6b), achieving better consistency between simulation ($W_{\text{simul}}^{\text{MuSiCal}}$) and data ($W_{\text{data}}$) in *in silico* validation (Fig. 7a). The assignments of other SBS and ID signatures are also much improved with MuSiCal, as indicated by the overall reduced bar heights in Fig. 7a and Extended Data Fig. 17 and the smaller cosine distances between W_{simul} and W_{data} in Fig. 7b. Of note, SBS5 appears to be over-assigned by both MuSiCal and PCAWG (blue bars in Fig. 7a), suggesting that the SBS5 spectrum in its current form may be inaccurate, or that it is in fact a composite signature corresponding to multiple mutational processes operative in different tumor types.

Previous studies [14] and our own analysis (Fig. 6c) suggest that SBS40 is correlated with ID5, characterized by 1-bp T deletions at 1 ∼ 5-bp mononucleotide T repeats (Fig. 6a). If SBS40 is truly a ubiquitous signature, this

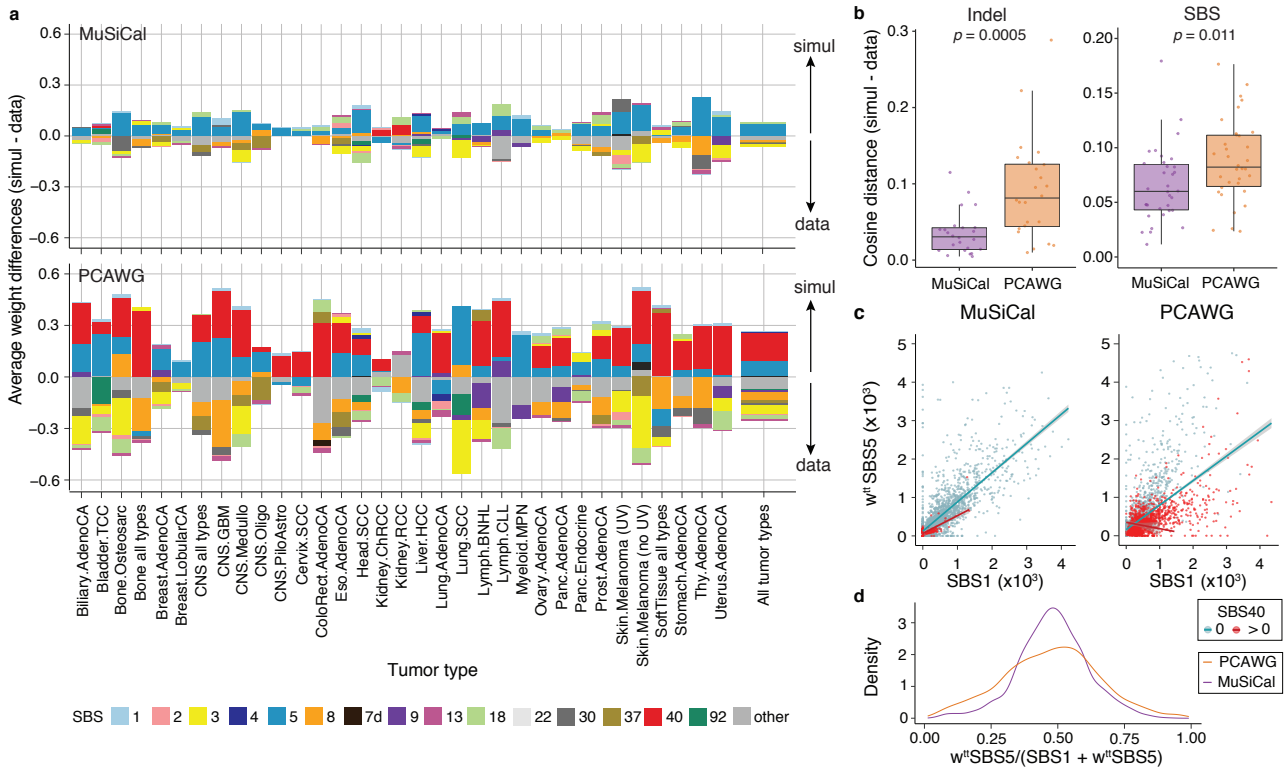


Fig. 7 MuSiCal improves the accuracy of signature assignments for PCAWG data and specifically solves the over-assignment problem of SBS40. a. Differences between the *de novo* signatures extracted from simulation (W_{simul}) and those from data (W_{data}) in terms of NNLS weights obtained by matching the corresponding *de novo* signatures to assigned signatures. The differences are calculated in the same way as in Fig. 5f, except that the average differences across *de novo* signatures from each tumor type are plotted. Positive (excess in simulation, indicating over-assignment) and negative (excess in data, indicating under-assignment) differences are plotted separately in opposite directions in a cumulative manner. PCAWG assignments show widespread over-assignment of SBS40 across diverse tumor types (red bars), while MuSiCal assignments demonstrate much improved consistency between simulation and data, as indicated by the overall reduced bar heights. **b.** Mean cosine distances between W_{simul} and W_{data} for ID (left) and SBS (right) signature assignments across different tumor types. MuSiCal assignments demonstrate improved consistency between simulation and data, as indicated by the smaller cosine distances overall. *p*-values are calculated with t-tests. **c.** Correlation between exposures of SBS1 and SBS5 for MuSiCal (left) and PCAWG (right) assignments. Samples with zero and nonzero exposures of SBS40 are plotted separately. To account for different accumulation rates of the two clock-like signatures in different tumor types, the SBS5 exposure is multiplied by a normalization factor w^{tt}, corresponding to the ratio between average SBS1 and SBS5 exposures for a given tumor type (tt) (Extended Data Fig. 19a). **d.** Distribution of the fraction of weighted SBS5 exposure over the sum of SBS1 and weighted SBS5 exposures. MuSiCal assignments demonstrate improved correlation between SBS1 and SBS5, as indicated by the tighter distribution.

correlation should be observed in most tumor types. However, a strong correlation between SBS40 and ID5 exists only in kidney and bladder cancers (for both MuSiCal and PCAWG assignments) and not in the other tumor types (for PCAWG assignments) (Extended Data Fig. 18). Therefore, the assignment of SBS40 in the other tumor types by PCAWG is unlikely to be reliable.

The impact of the over-assignment of SBS40 cannot be overstated, as it also results in widespread under-assignment of other similar or overlapping signatures. For example, in the results obtained by the PCAWG consortium, SBS8 (unknown etiology) is under-assigned in multiple tumor types where SBS40 is over-assigned (Fig. 7a, Extended Data Fig. 14). Similar under-assignments can be observed for SBS18 (damage by reactive oxygen species), 34 (unknown etiology), and 41 (unknown etiology) (Fig. 7a, Extended Data Fig. 14). To further demonstrate that the over-assignment of SBS40 can confound the interpretation of other signatures, we investigate the correlation between SBS1 and SBS5. Both signatures are referred to as clock-like signatures and correlate with patients' age in most cancers and normal cells [14, 33, 34, 62]. We thus expect to observe a correlation between the exposures of SBS1 and 5, although potentially with different slopes in different tumor types. When these tumor type-specific accumulation rates are factored out by multiplying the SBS5 exposure with a normalization factor corresponding to the ratio of average SBS1 and SBS5 exposures for a given tumor type (Extended Data Fig. 19), a constant slope should be observed in the pan-cancer setting. However, signature assignments from the PCAWG consortium show distinct SBS5/SBS1 slopes for samples with and without SBS40 (Fig. 7c, d). In particular, the normalized SBS5/SBS1 ratio is consistently lower in SBS40-positive samples, suggesting that a proportion of mutations caused by SBS5 are mis-attributed to SBS40 in these samples (Fig. 7c, Extended Data Fig. 19c). By comparison, MuSiCal-derived signature

assignments show consistent SBS5/SBS1 slopes that are not confounded by SBS40 (Fig. 7c, d, Extended Data Fig. 19c).

3 Discussion

Mutational signature analysis has become routine in cancer genome analysis [63]. In this work, we developed MuSiCal that utilizes novel algorithms to improve the performance of both signature discovery and assignment. For signature discovery, MuSiCal employs mvNMF; for signature assignment, all three innovations – mvNMF, likelihood-based sparse NNLS, and *in silico* validation via data-driven simulations – are pivotal. Importantly, the data-driven simulations provide a systematic method for optimizing parameters in post-discovery steps and for revealing potential issues in final signature assignments. Without such a validation module, widespread signature misassignments by previous tools have gone unchecked. In fact, the application of data-driven simulations is not limited to MuSiCal or MuSiCal-derived signature assignments. The same framework can be used to identify issues with signature assignments and tune parameters for other signature analysis tools.

The number of whole genomes of cancer and other diseases continues to grow rapidly, especially by consortium projects such as Genomics England's 100,000 Genomes Project [64] and Hartwig Medical Foundation's metastatic tumor project [65]. Applications of MuSiCal to these datasets will further refine the set of mutational signatures and facilitate comparison of signatures from different contexts, such as tumor types, metastatic status, and treatments received. To handle these datasets, the scalability of MuSiCal could be improved. The parameter optimization step based on data-driven simulations is especially expensive computationally, as *de novo* signature discovery needs to be repeated for each grid point. More efficient approaches, such as a gradient-based search in the parameter space, may be necessary.

Mutational signature catalogs (such as those from COSMIC) serve as a crucial resource for the interpretation of mutational spectra. Our reanalysis of PCAWG data with MuSiCal provides a substantial revision and expansion of the catalog of ID signatures. For SBS signatures, our reanalysis demonstrates that there are still new signatures to be discovered even from existing data and that the spectra of some known signatures – such as SBS40 (Extended Data Fig. 13) – need to be revised. Our analysis is based on the standard signature definitions, i.e., the 96-channel trinucleotide signatures for SBSs and the 83-channel signatures for IDs. Application of MuSiCal to alternative signature definitions should further disentangle mutational processes and help gain insights into signatures with unknown etiologies. In particular, the standard definition of ID signatures with 83 channels lacks more detailed indel characteristics, such as affected nucleotides of > 1-bp indels and nucleotide context of flanking sequences beyond repetitive and microhomology information (e.g., the specific motif as observed for TOP1 signature [58]). As a result, some of the ID signatures may correspond to multiple mutational processes. This redundancy may underlie the apparent lack of tissue specificity of ID signature assignments compared to SBS signature assignments (Fig. 6b). Signatures of other mutation types, including copy number variations and structural variations, are less well studied [15, 66–68]. We expect that MuSiCal will facilitate future developments of these signatures when their proper definitions are developed.

Supplementary information. Supplementary Tables 1-4 are available.

Acknowledgments. This work was supported by a grant from the Ludwig Center at Harvard and a grant by the Cancer Research UK Grand Challenge and the Mark Foundation to the SPECIFICANCER team. V.L. was supported by the Swedish Research Council (2020-00583). We would like to thank D. Glodzik and B. Geiger for careful reading of the manuscript.

4 Online Methods

4.1 Data and code availability

PCAWG data, including mutation count matrices of SBSs and IDs, as well as exposure matrices obtained by [14], were downloaded from https://www.synapse.org/#!Synapse:syn11726601/files/. Additional PCAWG data were downloaded from https://dcc.icgc.org/releases/PCAWG. All PCAWG data used in this paper are in the open tier. COSMIC signatures were downloaded from https://cancer.sanger.ac.uk/signatures/. SigProfilerExtractor was downloaded from https://github.com/AlexandrovLab/SigProfilerExtractor. sigLASSO was downloaded from https://github.com/gersteinlab/siglasso. MuSiCal is implemented in Python and available at https://github.com/parklab/MuSiCal.

4.2 The MuSiCal algorithm

The input of MuSiCal is a p by n matrix X of mutation counts, where p denotes the number of features (e.g., $p = 96$ for SBS signatures with trinucleotide contexts) and n the number of samples.

4.2.1 *De novo* signature discovery

In *de novo* signature discovery, MuSiCal decomposes X into the product of a p by r signature matrix W and a r by n exposure matrix H, where r denotes the number of signatures.

Minimum-volume NMF. Standard NMF tries to factorize X as a product of W and H by minimizing the reconstruction error $L = D(X|WH)$, where D denotes a distance measure, such as the Frobenius norm or the generalized Kullback-Leibler (KL) divergence. However, any invertible matrix Q provides an equivalent pair of solutions $(WQ, Q^{-1}H)$ as long as the resulting matrices are non-negative, resulting in NMF solutions being non-unique. MuSiCal employs minimum-volume NMF (mvNMF) to solve this non-uniqueness problem [43–46]. In mvNMF, the following optimization problem is solved [46],

$$\min_{W\geq 0, H\geq 0} \left(D(X|WH) + \lambda \log\det(W^T W + \delta I) \right), \tag{1}$$

where λ is the penalty parameter controlling the strength of volume regularization, δ a positive constant preventing $\log\det(W^T W)$ from going to $-\infty$, and I the identity matrix. In MuSiCal, we use KL divergence for the distance measure and set δ to 1. We follow the algorithm in [46] to solve the optimization problem above. Specifically, W and H are randomly initialized and alternately updated while keeping the other one fixed. When W is fixed, the problem is equivalent to standard NMF, and thus the well-known multiplicative update is used to update H [42]. When H is fixed, a revised multiplicative update is used to update W, combined with a backtracking line-search procedure to guarantee that columns of W are L_1 normalized [46]. The updates are iterated until the relative change of objective falls below a tolerance threshold.

Selecting the regularization parameter. In mvNMF, the regularization parameter λ needs to be carefully chosen for optimal performance. To make λ less dependent on the initialization and thus more generalizable across multiple runs, we set $\lambda = \tilde{\lambda}\frac{D(X|W_{\text{ini}}H_{\text{ini}})}{\log\det(W_{\text{ini}}^T W_{\text{ini}}+\delta I)}$, where W_{ini} and H_{ini} are initializations for W and H, respectively, and tune $\tilde{\lambda}$ instead [45]. To tune $\tilde{\lambda}$, mvNMF is run for a logarithmic grid (by default, from 10^{-10} to 2) of $\tilde{\lambda}$ values separately. Then, the solutions for different $\tilde{\lambda}$ values are compared to the one for the smallest $\tilde{\lambda}$ (by default, 10^{-10}), which is nearly equivalent to the standard NMF solution without any regularization. We reason that, as long as the solution is not significantly worse than the NMF solution in reconstructing X, the strongest volume regularization should be favored. Therefore, we choose the largest $\tilde{\lambda}$ value such that the reconstruction error $D(X|WH)$ is not significantly worse when compared to the solution with the smallest $\tilde{\lambda}$. Specifically for comparing reconstruction errors, two statistical tests are jointly used. First, the Mann-Whitney U test is used to compare sample-wise reconstruction errors. Second, to better capture potential outliers that receive exceptionally large errors, a test for differences in distribution tails is used [69]. A solution is said to be significantly worse if either of the two tests gives a p-value smaller than 0.05.

Solution filtering and clustering. For a fixed choice of r (number of signatures), mvNMF is run multiple times (by default, 20) with different initializations. Of note, $\tilde{\lambda}$ tuning is performed only once, and the selected $\tilde{\lambda}$ is used for the rest of the mvNMF runs, as we have observed that the optimal $\tilde{\lambda}$ is stable with a fixed r and X, even when X is bootstrapped before each run. To get rid of occasional bad solutions potentially due to failed convergence, solutions with reconstruction errors greater than 5 median absolute deviations (MAD) above the median are filtered out. *De*

novo signatures (i.e., columns of W) of the remaining solutions are clustered via hierarchical clustering, and cluster means are taken as the final solutions for *de novo* signatures.

Selecting the number of signatures. In practice, the true number of signatures r is not known *a priori* and needs to be inferred from the data itself. MuSiCal employs a novel method for automatic selection of r (Extended Data Fig. 6). After running mvNMF multiple times with different initializations for each candidate r and filtering out potential bad solutions (as described above), previous methods usually proceed by clustering the signatures from solutions with the same r into r clusters, and choosing the best r such that the reconstruction error is reasonably small and the solutions are reasonably stable, as quantified by the silhouette score of the clustering [7, 30, 53]. However, the exact choice of r is either manual or heavily dependent on predefined thresholds. More importantly, forcing the number of clusters to be equal to r seems artificial since multiple mvNMF/NMF runs may in fact produce different signatures, especially when r is misspecified (see, for example, Extended Data Fig. 6c). We therefore propose to cluster the signatures from solutions with the same r into k clusters, allowing $k \neq r$, where the optimal k is determined by the well-established gap statistic [54, 55]. We reason that, at the true underlying r, solutions from multiple mvNMF/NMF runs should be stable and comprise the same signatures, leading to $k = r$. By contrast, when r is over-specified, mvNMF/NMF should start discovering redundant or split signatures, resulting in $k \neq r$ (most likely, $k < r$) (see, for example, Extended Data Fig. 6g, h). We thus select the largest r among those with $k = r$ as the inferred number of signatures (Extended Data Fig. 6b). This method is parameter-free and more robust. In practice, $k = r$ can happen accidentally when r is over-specified. We therefore additionally check the per-signature silhouette scores to ensure that these occasional cases are not selected. By default, the largest r with $k = r$ as well as a mean silhouette score ≥ 0.7 and a minimum silhouette score ≥ 0.2 is selected.

4.2.2 Matching and refitting

Following mvNMF-based *de novo* signature discovery, matching and refitting are performed. In matching, each *de novo* signature is matched to a known catalog, e.g., from COSMIC, and decomposed as a non-negative mixture of potentially multiple signatures from the catalog. In refitting, the original data X is refit against matched signatures from the catalog to recalculate the corresponding exposures. MuSiCal uses a novel multinomial likelihood-based sparse NNLS algorithm for both matching and refitting.

Multinomial likelihood-based sparse NNLS. We describe a generic problem of fitting a sample spectrum, denoted as a p-dimensional column vector x of mutation counts (e.g., a column of X), against a set of signatures, represented as columns of the p by r matrix W. W can be the set of matched signatures from the catalog if refitting is performed after *de novo* discovery and matching, or it can be any user-specified signatures if refitting is performed as a standalone task.

In standard NNLS, the r-dimensional column vector h of exposures is estimated by solving the following optimization problem,

$$\min_{h \geq 0} \|x - Wh\|_2^2, \tag{2}$$

where $\|\cdot\|_2$ denotes the L_2 norm. NNLS lacks explicit sparsity constraints and leads to the over-assignment problem. We thus propose a novel multinomial likelihood-based sparse NNLS algorithm (Extended Data Fig. 8a). In this algorithm, we first perform NNLS to obtain an initial exposure estimate $h^{(0)}$ and then refine the exposure vector in a bidirectional stepwise manner.

Let $h^{(m)}$ be the exposure estimate at the m-th iteration. The multinomial log likelihood of observing x according to $h^{(m)}$ is

$$\log L^{(m)}(h^{(m)}|x) = \sum_{i=1}^{p} x_i \log w_i^{(m)}, \tag{3}$$

where $w^{(m)} = \frac{Wh^{(m)}}{\sum_{i=1}^{p}(Wh^{(m)})_i}$ is the normalized probability vector of the composite signature corresponding to $h^{(m)}$, and the constant combinatorial factor is dropped since it is independent of $h^{(m)}$ and depends only on the fixed x. Let $R^{(m)} = \{j = 1, 2, \ldots, r | h_j^{(m)} > 0\}$ be the set of indices of nonzero exposures in $h^{(m)}$, i.e., the support of $h^{(m)}$ or the set of indices corresponding to signatures believed to be active at the m-th iteration. Let $\bar{R}^{(m)} = \{j = 1, 2, \ldots, r | h_j^{(m)} = 0\}$ be the complement of $R^{(m)}$.

During the backward step, a currently active signature is excluded if removing it does not significantly reduce the multinomial likelihood of the observed mutation count vector x. Specifically, we attempt to exclude each active signature in turn and calculate the revised multinomial log likelihood $\log L^{(m)(-j)}(h^{(m)(-j)}|x), \forall j \in R^{(m)}$, where $h^{(m)(-j)}$ is the exposure vector returned by NNLS when x is fit against the signatures with indices $R^{(m)} \setminus \{j\}$, while all other signatures corresponding to $\bar{R}^{(m)} \cup \{j\}$ are forced to have zero exposures. Then, the revised log likelihoods

$\log L^{(m)(-j)}$ are compared to the original $\log L^{(m)}$; the signature $\hat{j}$ with the smallest drop in log likelihood when its exposure is additionally forced to be zero is selected; signature $\hat{j}$ is excluded if the drop in log likelihood is smaller than a predefined positive threshold $\epsilon > 0$; otherwise, $h^{(m)}$ is kept unchanged. That is,

$$\begin{aligned} h^{(m+1)} &= h^{(m)(-\hat{j})}, \text{if } \log L^{(m)} - \log L^{(m)(-\hat{j})} < \epsilon, \\ h^{(m+1)} &= h^{(m)}, \text{if } \log L^{(m)} - \log L^{(m)(-\hat{j})} \geq \epsilon, \\ \hat{j} &= \arg\min_{j \in R^{(m)}} (\log L^{(m)} - \log L^{(m)(-j)}). \end{aligned} \tag{4}$$

Similarly, during the forward step, a currently inactive signature is included if adding it significantly increases the multinomial likelihood of observing x,

$$\begin{aligned} h^{(m+1)} &= h^{(m)(+\hat{j})}, \text{if } \log L^{(m)(+\hat{j})} - \log L^{(m)} > \epsilon, \\ h^{(m+1)} &= h^{(m)}, \text{if } \log L^{(m)(+\hat{j})} - \log L^{(m)} \leq \epsilon, \\ \hat{j} &= \arg\max_{j \in \bar{R}^{(m)}} (\log L^{(m)(+j)} - \log L^{(m)}), \end{aligned} \tag{5}$$

where $h^{(m)(+j)}$ is the exposure vector returned by NNLS when x is fit against the signatures with indices $R^{(m)} \cup \{j\}, \forall j \in \bar{R}^{(m)}$, while all other signatures $\bar{R}^{(m)} \setminus \{j\}$ are forced to have zero exposures.

The backward and forward steps are alternately iterated until no signature is removed or added. The same likelihood threshold ϵ is used for both steps to reduce the number of parameters and ensure that successive backward and forward steps do not remove or add the same signature. We have observed in simulation studies that even when different thresholds are allowed for backward and forward steps, the global optimum is usually achieved with the same threshold.

Of note, the log likelihood in Equation (3) is divided by the total mutation count $\sum_{i=1}^{p} x_i$ in practice, leading to a per-mutation log likelihood. This normalization makes the likelihood threshold ϵ independent of the total mutation count (thus more generalizable) and turns out to be crucial for adapting the algorithm for matching (see below).

Adapting the likelihood-based sparse NNLS for matching. Matching and refitting are essentially the same mathematical problem, except that in matching, x is an L_1-normalized probability vector corresponding to the *de novo* signature instead of an integer vector of mutation counts. The multinomial log likelihood of observing a probability vector x as in Equation (3) may seem ill-defined at first glance. However, observe that Equation (3) is equivalent to the KL divergence between x and $w^{(m)}$ if the constant term independent of $h^{(m)}$ is dropped. Note that both x and $w^{(m)}$ are well-defined probability vectors – $w^{(m)}$ is L_1-normalized by definition; x is L_1-normalized in matching since it is in fact a *de novo* signature; and in refitting, the normalization by total mutation count effectively turns x into a probability vector. Therefore, the multinomial likelihood-based sparse NNLS can be alternatively interpreted as a bidirectional stepwise NNLS based on KL divergence. In fact, this correspondence is natural given the keen relation between likelihood and KL divergence. Consequently, the same algorithm can be directly used for matching without any modifications.

4.2.3 Data-driven parameter optimization and *in silico* validation

MuSiCal employs a novel data-driven approach for optimizing parameters involved in post-processing steps after *de novo* signature discovery, as well as validating the consistency between the final signature assignment and the original data *in silico* (Fig. 5a).

Let X_{data} be the original data matrix of mutation counts, and $(W_{\text{data}}, H_{\text{data}})$ the corresponding *de novo* discovery results. After matching and refitting with properly chosen parameters (see below), the final signature assignments $(W_{\text{s}}, H_{\text{s}})$ are obtained, where s stands for sparse. In order to assess the quality of $(W_{\text{s}}, H_{\text{s}})$, we simulate a synthetic mutation count matrix X_{simul} from $(W_{\text{s}}, H_{\text{s}})$ through multinomial sampling, and perform *de novo* signature discovery from X_{simul} with exactly the same settings as used for X_{data}. In particular, the selected number of signatures r and mvNMF regularization parameter $\tilde{\lambda}$ are kept the same. Let $(W_{\text{simul}}, H_{\text{simul}})$ be the *de novo* discovery results from X_{simul}. The distances between $(W_{\text{simul}}, H_{\text{simul}})$ and $(W_{\text{data}}, H_{\text{data}})$ are then used to quantify the consistency between $(W_{\text{s}}, H_{\text{s}})$ and X_{data}. In more detail, a one-to-one correspondence between signatures in W_{simul} and W_{data} is determined to minimize the mean cosine distance, and the resulting mean cosine distance is used to quantify the overall quality of $(W_{\text{s}}, H_{\text{s}})$.

MuSiCal also applies the same scheme for optimizing post-processing parameters (Extended Data Fig. 9). Two likelihood thresholds, corresponding to the likelihood-based sparse NNLS used for matching and refitting, respectively,

are optimized through a two-dimensional grid search. Specifically, the threshold pair that minimizes the mean cosine distance between W_{simul} and W_{data} (as described above) is selected as a candidate solution. Additional threshold pairs are further included in the candidate list if the corresponding difference between W_{simul} and W_{data} is not significantly larger (i.e., if both the Mann-Whitney U test and the test for differences in distribution tails [69] result in p-values > 0.05 for the element-wise L_1 error). Then, among these candidate threshold pairs, the one with the smallest number of assigned signatures is selected, and if there are multiple such pairs, the one with the largest thresholds (i.e., the sparsest signature assignment) is selected. Alternatively, to reduce computation time, two separate one-dimensional grid searches can be performed (Extended Data Fig. 9): the matching threshold is first optimized while the refitting threshold is fixed to a small value (0.0001), and the refitting threshold is subsequently optimized while the matching threshold is fixed to its optimal value.

It is worth pointing out that the application of this data-driven approach is not limited to post-processing steps in MuSiCal or MuSiCal-derived signature assignments. Indeed, the quality of signature assignments obtained by the PCAWG consortium can be assessed within the same framework, and potential issues can be detected (e.g., see Fig. 5). Parameters used in post-processing steps of other signature analysis tools can also be optimized with the same approach, although a high-dimensional grid search can be prohibitive when many parameters are involved.

4.2.4 Preprocessing

MuSiCal provides two complementary methods for preprocessing in order to improve the sensitivity of *de novo* signature discovery (Extended Data Fig. 10).

Automatic cohort stratification. In automatic cohort stratification, samples in the input matrix X are clustered into potential subsets with distinct signature compositions, e.g., MMRD vs. MMRP, samples with vs. without aristolochic acid exposure, etc. (Extended Data Fig. 10a-c). Such stratifications may benefit *de novo* signature discovery by allowing subtle signatures within each subset to be discovered cleanly. Specifically, X is subject to hierarchical clustering, where the optimal number of clusters k is determined by the gap statistic [54, 55]. Compared to other metrics (such as the silhouette score) that are ill-defined when the number of clusters is 1, the gap statistic handles $k = 1$ gracefully and is able to select $k = 1$ when there is in fact no subset structure within the dataset. When $k > 1$ is selected, *de novo* signature discovery is run on the corresponding subsets of samples separately. Alternatively, the clustering is performed on the *de novo* exposure matrix H after an initial *de novo* signature discovery on X with all samples. We have observed that the latter approach of clustering H better captures inherent subset structures within the dataset and is thus recommended. Of note, cohort stratification can lead to subsets with small numbers of samples, reducing the power of signature discovery. Therefore, the suggested stratification needs to be investigated in a case-specific manner in order to maximize the benefit for signature discovery (see below for detailed approach taken for reanalysis of PCAWG data).

Outlier removal based on the Gini coefficient. When the dataset contains a small number of samples with strong exposures of signatures that are not present in other samples, these strong signatures can reduce the sensitivity of detecting other more prevalent signatures, or introduce biases in them, during *de novo* signature discovery. We therefore detect such outlier samples and remove them to increase sensitivity and accuracy of signature discovery (Extended Data Fig. 10d-f). Specifically, after an initial signature discovery with the entire dataset, we investigate the distribution of *de novo* exposures in H with the Gini coefficient for each *de novo* signature. The Gini coefficient quantifies the inequality among the per-sample exposures – a small Gini coefficient is expected for a homogeneous dataset with similar signature compositions, while strong outliers are indicated by a large Gini coefficient. We thus first select signatures for which the exposures have a Gini coefficient greater than a baseline threshold (by default 0.65). Then, for each selected signature, samples are inspected one by one in descending order of the corresponding relative exposure, and a sample is excluded if removing it results in a decrease in the Gini coefficient greater than a threshold δ_{gini} (by default, 0.005).

4.3 Simulation studies

4.3.1 Simulation studies comparing mvNMF and NMF for *de novo* signature discovery

To compare mvNMF and NMF for *de novo* signature discovery (Fig. 2c, Extended Data Fig. 3), synthetic datasets were simulated from tumor type-specific SBS signatures and Dirichlet-distributed exposures. In more detail, SBS signatures specifically present in 32 PCAWG tumor types were read out from Fig. 3 in [14]. Random exposures were generated from symmetric Dirichlet distributions with a concentration parameter of $\alpha = 0.1$, which is a representative value according to real exposure matrices obtained by the PCAWG consortium [14] (Extended Data Fig. 4). Mutation count matrices were then simulated through multinomial sampling, and 20 independent simulations were performed

for each tumor type. In order to focus on algorithmic differences between NMF and mvNMF, each synthetic dataset was allowed to contain 200 samples and on average 5,000 mutations per sample (Poisson-distributed), such that there was enough power to discover each signature involved. The number of signatures was also assumed to be known. NMF and mvNMF were then applied to derive *de novo* signatures, and the same random initialization was used for both algorithms for a given dataset to facilitate comparison. For NMF, the implementation within MuSiCal according to the multiplicative update algorithm with KL divergence was used [42].

4.3.2 Simulation studies benchmarking MuSiCal and SigProfilerExtractor for *de novo* signature discovery

To benchmark the performance of MuSiCal and SigProfilerExtractor for *de novo* signature discovery (Fig. 3, Extended Data Fig. 7), more realistic synthetic datasets were simulated to capture the complexity present in real-life data. Specifically, tumor type-specific SBS signatures were combined with real exposure matrices obtained by the PCAWG consortium [14] to simulate mutation count matrices through multinomial sampling. 10 independent simulations were performed for each of the 25 tumor types with at least 20 samples remaining after removing hypermutated samples with signatures associated with MMRD or DNA polymerase mutations (SBS6, 9, 10, 14, 15, 20, 21, 26, and 44). Samples with MMRD or DNA polymerase mutations are usually analyzed as a separate group for improved sensitivity of signature discovery [14] and thus excluded. MuSiCal and SigProfilerExtractor were then applied to identify *de novo* signatures. For SigProfilerExtractor, version 1.1.3 was used with default parameters (`-i random -b True -nr 100 -min 10000 -max 1000000 -conv 10000 -tol 1e-15`). Input normalization was turned off (`-nx none`) for Fig. 3 and Extended Data Fig. 7a, b, and turned on (`-nx gmm`) for Extended Data Fig. 7c, d. For MuSiCal, the preprocessing module was skipped for both Fig. 3 and Extended Data Fig. 7. In Fig. 3 and Extended Data Fig. 7c, d, the Decomposition module from SigProfilerExtractor was used with default parameters (`nnls_add_penalty = 0.05, nnls_remove_penalty = 0.01, initial_remove_penalty = 0.05, de_novo_fit_penalty = 0.02, connected_sigs = True`) to match both MuSiCal- and SigProfilerExtractor-derived *de novo* signatures to the COSMIC catalog. In Extended Data Fig. 7a, b, the likelihood-based sparse NNLS from MuSiCal was used to match both MuSiCal- and SigProfilerExtractor-derived *de novo* signatures to the COSMIC catalog, and the likelihood threshold was separately adjusted for each solution to achieve the optimal F1 score, in order to minimize the impact of this matching step on evaluating *de novo* signatures.

4.3.3 Simulation studies benchmarking different refitting algorithms

To benchmark the performance of different algorithms for refitting (Fig. 4c, d), synthetic datasets were simulated from tumor type-specific SBS signatures and Dirichlet-distributed ($\alpha = 0.1$) exposures, similar to those in Fig. 2c and Extended Data Fig. 3. Specifically, 100 samples were simulated for each of the 32 tumor types and combined together as a single dataset. Each sample contained 5000 mutations on average (Poisson-distributed), and 10 independent simulations were performed. To further introduce sparsity in the exposure matrices, signatures with relative exposures < 0.01 as sampled from the Dirichlet distribution were set to have zero exposures. Four different algorithms were then applied to refit these synthetic samples against the COSMIC catalog. For likelihood-based sparse NNLS and thresholded NNLS (also implemented in MuSiCal), the optimal threshold was chosen through a grid search to achieve the maximum correct support discovery rate when averaged across the 10 independent simulations. For sigLASSO, version 1.1 was used with default parameters and flat priors. For SigProfilerExtractor, the Decomposition module from version 1.1.3 was used with default parameters, except that `connected_sigs = False` was set to avoid the use of empirical rules.

4.4 Reanalysis of PCAWG data

4.4.1 SBS signatures

***De novo* signature discovery.** Samples with MMRD and/or POLE-exo mutations were isolated and analyzed separately in a tissue-independent manner. In order to identify these samples, we inspected the exposures of SBS signatures known to be associated with MMRD and/or POLE-exo mutations (MMRD: SBS6, 14, 15, 20, 21, 26, and 44; POLE-exo: SBS10a-d, and 28), together with the microsatellite mutation rate as obtained from [70] (Extended Data Fig. 20a, b). MMRD/POLE tumors showed a clear separation from the MMRP ones based on these metrics and were subsequently identified (40 tumors with MMRD, 9 with POLE-exo mutations, and 1 with concurrent MMRD and POLE-exo mutations) (Extended Data Fig. 20a, b). Specifically for this identification step, exposures of relevant signatures were determined simply by standard NNLS using all signatures in the COSMIC catalog. Then, *de novo* signature discovery followed by matching and refitting was performed to determine the final signature assignments.

During the initial signature discovery, we found that signatures also present in MMRP samples (such as SBS1, 2, 4, 7a, 7b, 17b, and 30) were mixed with signatures specific to MMRD/POLE-exo mutant tumors. To facilitate the accurate extraction of these MMRD/POLE-specific signatures, we included spectra of the discovered MMRP signatures (scaled by the median SBS count in this cohort) as additional samples and performed *de novo* signature discovery again.

The remaining samples were grouped into 37 tumor types as defined by the PCAWG consortium. *De novo* signature discovery was performed separately for each of the 26 tumor types with at least 20 samples (100 on average). The remaining 11 tumor types had too few samples for independent signature discoveries. For example, there were only 2 samples for Cervix.AdenoCA, 3 for Breast.DCIS, and 4 for Myeloid.MDS. These 11 tumor types were of 6 different tissue origins (breast, cervix, myeloid, bone, soft tissue, and CNS) and thus combined with tumor types with the same tissue origin, respectively, for *de novo* signature discovery. For example, samples from Breast.DCIS and Breast.LobularCA were combined with those from Breast.AdenoCA, and the resulting larger cohort was used for signature discovery. Of note, the combined cohorts were only used to inform the 11 tumor types with a small number of samples, while the 26 tumor types with a sufficient number of samples were still analyzed separately.

In addition to the large variation in number of samples per tumor type, the number of mutations per sample also varied greatly within each tumor type. In particular, the presence of hypermutated samples can significantly influence the signature discovery results, as a single highly mutated sample can dominate the objective function optimized by mvNMF and thus be picked up as a distinct signature even when it is actually a composite spectrum. We therefore performed L_1 normalization for columns of the input matrix X before running *de novo* signature discovery. We also excluded samples with less than 500 SBSs, as samples with few mutations can cause biases after L_1 normalization. As a result, 5.6% of samples were removed in total, and less than 21% of samples were removed from all tumor types except CNS.PiloAstro, in which 91% of samples had less than 500 SBSs. CNS.PiloAstro was thus combined with other samples with CNS origin for *de novo* signature discovery, resulting in 25 tumor types processed through tumor type-specific analysis and 12 tumor types processed through tissue-of-origin-specific analysis in the end.

For each tumor type, the optimal $\tilde{\lambda}$ for mvNMF varied between $0.01 \sim 0.5$ with an average of 0.05. The selected number of signatures r varied between $2 \sim 15$ with an average of 6. The dependence of $\tilde{\lambda}$ and r on the number of samples and SBSs is shown in Extended Data Fig. 20c.

Implementation of preprocessing. After an initial step of *de novo* signature discovery within each tumor type, the Gini-based outlier removal was applied, removing on average 4.1% of samples. Next, we stratified samples into subsets with distinct signature compositions with hierarchical clustering. Note that we separated the MMRD and POLE-exo mutant samples before this stage, although preprocessing can be used to stratify these samples if they had not been isolated beforehand (Extended Data Fig. 10a-c).

Matching, refitting and validation. Matching of *de novo* signatures to the COSMIC catalog and refitting were performed with likelihood-based sparse NNLS. The two likelihood thresholds were optimized with a grid search using data-driven simulations as in the *in silico* validation/optimization module of MuSiCal. In the matching step, three signatures – SBS25, 39, and 86 – were excluded. These signatures contributed small weights (< 0.3) in matching and were usually identified only when other overlapping signatures were present (SBS8, 22, and 35 for SBS25 with overlapping T>A mutations, SBS3 and 13 for SBS39 and 86 with overlapping C>G mutations). Excluding these signatures had minimal effect, resulting in a difference of ~ 0.02 in the cosine reconstruction errors of *de novo* signatures. Further, signatures associated with MMRD and/or POLE-exo mutations were only used in matching when analyzing samples with MMRD and/or POLE-exo mutations, and UV-related signatures (SBS7a-d) were limited to the melanoma cohort. Finally, an additional cleaning step was performed after matching to avoid flat signatures being picked up due to tiny backgrounds present in the *de novo* signatures. Specifically, a matched signature was considered to be a consequence of overfitting to the background if (1) the sum of its probabilities at the 96 trinucleotide types where it contributed at least 1% was less than 15% and (2) at the same time its maximum contribution to any of the trinucleotide types was five times smaller than the maximum probability of the *de novo* signature.

4.4.2 ID signatures

Similar to the analysis of SBS signatures, samples with MMRD and/or POLE-exo mutations were analyzed separately for ID signatures. However, a tumor type-specific *de novo* signature discovery for ID signatures was more challenging because of the lower number of IDs per sample. We therefore performed joint analysis of multiple tumor types to obtain well separated signatures. Specifically, we first ran *de novo* signature discovery with all samples having at least 100 IDs (n = 2241) combined together. We then performed hierarchical clustering with the obtained exposure matrix H, which resulted in two major clusters. Each tumor type was subsequently examined to determine which cluster contained the majority of samples from the tumor type. Accordingly, all tumor types were stratified into two

groups. *De novo* signature discovery was performed again on these two groups separately. The resulting signatures from these two groups were combined with those from the initial run with all tumor types to derive the updated ID signature catalog. Signature assignments were still performed for each tumor type separately, where tumor type-specific thresholds were used for the minimum number of IDs per sample to maximize the number of samples retained while reducing samples with low ID counts that deteriorate the result.

References

[1] Hollstein, M., Sidransky, D., Vogelstein, B. & Harris, C. C. p53 mutations in human cancers. Science **253** (5015), 49–53 (1991) .

[2] Hainaut, P. & Pfeifer, G. P. T transversions in lung cancers reflect the primary mutagenic signature of DNA-damage by tobacco smoke. Carcinogenesis **22** (3), 367–374 (2001) .

[3] Pfeifer, G. P., You, Y. H. & Besaratinia, A. Mutations induced by ultraviolet light. Mutat Res **571** (1-2), 19–31 (2005) .

[4] Macé, K. et al. Aflatoxin B1-induced DNA adduct formation and p53 mutations in CYP450-expressing human liver cell lines. Carcinogenesis **18** (7), 1291–1297 (1997) .

[5] Nedelko, T., Arlt, V. M., Phillips, D. H. & Hollstein, M. TP53 mutation signature supports involvement of aristolochic acid in the aetiology of endemic nephropathy-associated tumours. Int J Cancer **124** (4), 987–990 (2009) .

[6] Nik-Zainal, S. et al. Mutational Processes Molding the Genomes of 21 Breast Cancers. Cell **149** (5), 979–993 (2012) .

[7] 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 Reports **3** (1), 246–259 (2013) .

[8] Abbasi, A. & Alexandrov, L. B. Significance and limitations of the use of next-generation sequencing technologies for detecting mutational signatures. DNA Repair (Amst) **107**, 103200 (2021) .

[9] Koh, G., Zou, X. & Nik-Zainal, S. Mutational signatures: experimental design and analytical framework. Genome Biol **21** (1), 37 (2020) .

[10] Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet **15** (9), 585–598 (2014) .

[11] Koh, G., Degasperi, A., Zou, X., Momen, S. & Nik-Zainal, S. Mutational signatures: emerging concepts, caveats and clinical applications. Nat Rev Cancer **21** (10), 619–637 (2021) .

[12] Kim, Y. A. et al. Mutational Signatures: From Methods to Mechanisms. Annu Rev Biomed Data Sci **4**, 189–206 (2021) .

[13] Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. Nature **500** (7463), 415–421 (2013) .

[14] Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. Nature **578** (7793), 94–101 (2020) .

[15] Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. Nature **578** (7793), 112–121 (2020). Number: 7793 Publisher: Nature Publishing Group .

[16] Kucab, J. E. et al. A Compendium of Mutational Signatures of Environmental Agents. Cell **177** (4), 821–836.e16 (2019) .

[17] Zou, X. et al. Validating the concept of mutational signatures with isogenic cell models. Nature Communications **9** (1), 1744 (2018) .

[18] Zou, X. et al. A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. Nat Cancer **2** (6), 643–657 (2021) .

[19] Cook, J. H., Melloni, G. E. M., Gulhan, D. C., Park, P. J. & Haigis, K. M. The origins and genetic interactions of KRAS mutations are allele- and tissue-specific. Nat Commun **12** (1), 1808 (2021) .

[20] Muiños, F., Martínez-Jiménez, F., Pich, O., Gonzalez-Perez, A. & Lopez-Bigas, N. In silico saturation mutagenesis of cancer genes. Nature **596** (7872), 428–432 (2021) .

[21] Brady, S. W., Gout, A. M. & Zhang, J. Therapeutic and prognostic insights from the analysis of cancer mutational signatures. Trends Genet (2021) .

[22] Hoang, P. H., Cornish, A. J., Dobbins, S. E., Kaiser, M. & Houlston, R. S. Mutational processes contributing to the development of multiple myeloma. Blood Cancer J **9** (8), 60 (2019) .

[23] Secrier, M. et al. Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. Nat Genet **48** (10), 1131–1141 (2016) .

[24] Davies, H. et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat Med **23** (4), 517–525 (2017) .

[25] Staaf, J. et al. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. Nat Med **25** (10), 1526–1533 (2019) .

[26] Gulhan, D. C., Lee, J. J.-K., Melloni, G. E. M., Cortés-Ciriano, I. & Park, P. J. Detecting the mutational signature of homologous recombination deficiency in clinical samples. Nature Genetics **51** (5), 912 (2019) .

[27] Färkkilä, A. et al. Immunogenomic profiling determines responses to combined PARP and PD-1 inhibition in ovarian cancer. Nature Communications **11** (1), 1–13 (2020) .

[28] Maura, F. et al. A practical guide for mutational signature analysis in hematological malignancies. Nature Communications **10** (1), 2969 (2019). Number: 1 Publisher: Nature Publishing Group .

[29] Nguyen, L., W M Martens, J., Van Hoeck, A. & Cuppen, E. Pan-cancer landscape of homologous recombination deficiency. Nat Commun **11** (1), 5584 (2020) .

[30] Degasperi, A. et al. A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. Nat Cancer **1** (2), 249–263 (2020) .

[31] Thatikonda, V. et al. Comprehensive analysis of mutational signatures in pediatric cancers. bioRxiv (2021) .

[32] Guo, Q. et al. The mutational signatures of formalin fixation on the human genome. bioRxiv (2021) .

[33] Moore, L. et al. The mutational landscape of human somatic and germline cells. Nature **597** (7876), 381–386 (2021) .

[34] Moore, L. et al. The mutational landscape of normal human endometrial epithelium. Nature **580** (7805), 640–646 (2020) .

[35] Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole genome sequences. Nature **534** (7605), 47–54 (2016) .

[36] Sachs, N. et al. A Living Biobank of Breast Cancer Organoids Captures Disease Heterogeneity. Cell **172** (1-2), 373–386.e10 (2018) .

[37] Hoang, P. H. et al. Whole-genome sequencing of multiple myeloma reveals oncogenic pathways are targeted somatically through multiple mechanisms. Leukemia **32** (11), 2459–2470 (2018) .

[38] Le, D. T. et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science **357** (6349), 409–413 (2017) .

[39] Buisson, R., Lawrence, M. S., Benes, C. H. & Zou, L. APOBEC3A and APOBEC3B Activities Render Cancer Cells Susceptible to ATR Inhibition. Cancer Res **77** (17), 4567–4578 (2017) .

[40] Omichessan, H., Severi, G. & Perduca, V. Computational tools to detect signatures of mutational processes in DNA from tumours: A review and empirical comparison of performance. PLoS One **14** (9), e0221235 (2019) .

[41] Tate, J. G. et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res **47** (D1), D941–D947 (2019) .

[42] Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. Nature **401** (6755), 788–791 (1999) .

[43] Craig, M. D. Minimum-volume transforms for remotely sensed data. IEEE Transactions on Geoscience and Remote Sensing **32** (3), 542–552 (1994) .

[44] Miao, L. & Qi, H. Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. IEEE Transactions on Geoscience and Remote Sensing **45** (3), 765 (2007) .

[45] Ang, A. M. S. & Gillis, N. Algorithms and comparisons of nonnegative matrix factorizations with volume regularization for hyperspectral unmixing. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12** (12), 4843–4853 (2019) .

[46] Leplat, V., Gillis, N. & Ang, A. M. Blind audio source separation with minimum-volume beta-divergence NMF. IEEE Transactions on Signal Processing **68**, 3400–3410 (2020) .

[47] Campbell, P. J. et al. Pan-cancer analysis of whole genomes. Nature **578** (7793), 82–93 (2020). Number: 7793 Publisher: Nature Publishing Group .

[48] Huang, K., Sidiropoulos, N. D. & Swami, A. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. IEEE Transactions on Signal Processing **62** (1), 211–224 (2013) .

[49] Fu, X., Huang, K. & Sidiropoulos, N. D. On identifiability of nonnegative matrix factorization. IEEE Signal Processing Letters **25** (3), 328–332 (2018) .

[50] Fu, X., Huang, K., Sidiropoulos, N. D. & Ma, W.-K. Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications. IEEE Signal Process. Mag. **36** (2), 59–80 (2019) .

[51] Huang, K., Sidiropoulos, N. D. & Swami, A. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. IEEE Transactions on Signal Processing **62** (1), 211–224 (2013) .

[52] Seplyarskiy, V. B. et al. Population sequencing data reveal a compendium of mutational processes in the human germ line. Science **373** (6558), 1030–1035 (2021) .

[53] Islam, S. M. A. et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. bioRxiv (2021) .

[54] Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **63** (2), 411–423 (2001) .

[55] Mohajer, M., Englmeier, K.-H. & Schmid, V. J. A comparison of gap statistic definitions with and without logarithm function. arXiv preprint arXiv:1103.4767 (2011) .

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

[57] Li, S., Crawford, F. W. & Gerstein, M. B. Using sigLASSO to optimize cancer mutation signatures jointly with sampling likelihood. Nat Commun **11** (1), 3575 (2020) .

[58] Reijns, M. A. M. et al. Signatures of TOP1 transcription-associated mutagenesis in cancer and germline. Nature **602** (7898), 623–631 (2022) .

[59] Boot, A. et al. Recurrent mutations in topoisomerase IIα cause a previously undescribed mutator phenotype in human cancers. Proc Natl Acad Sci U S A **119** (4) (2022) .

[60] Luquette, L. J. et al. Ultraspecific somatic SNV and indel detection in single neurons using primary template-directed amplification. BioRxiv (2021) .

[61] Wojtowicz, D., Hoinka, J., Amgalan, B., Kim, Y. A. & Przytycka, T. M. RepairSig: Deconvolution of DNA damage and repair contributions to the mutational landscape of cancer. Cell Syst **12** (10), 994–1003 (2021) .

[62] Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. Nature Genetics **47** (12), 1402–1407 (2015) .

[63] Cortés-Ciriano, I., Gulhan, D. C., Lee, J. J., Melloni, G. E. M. & Park, P. J. Computational analysis of cancer genome sequencing data. Nat Rev Genet (2021) .

[64] Turnbull, C. et al. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. Bmj **361** (2018) .

[65] Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature **575** (7781), 210–216 (2019) .

[66] Macintyre, G. et al. Copy number signatures and mutational processes in ovarian carcinoma. Nat Genet **50** (9), 1262–1270 (2018) .

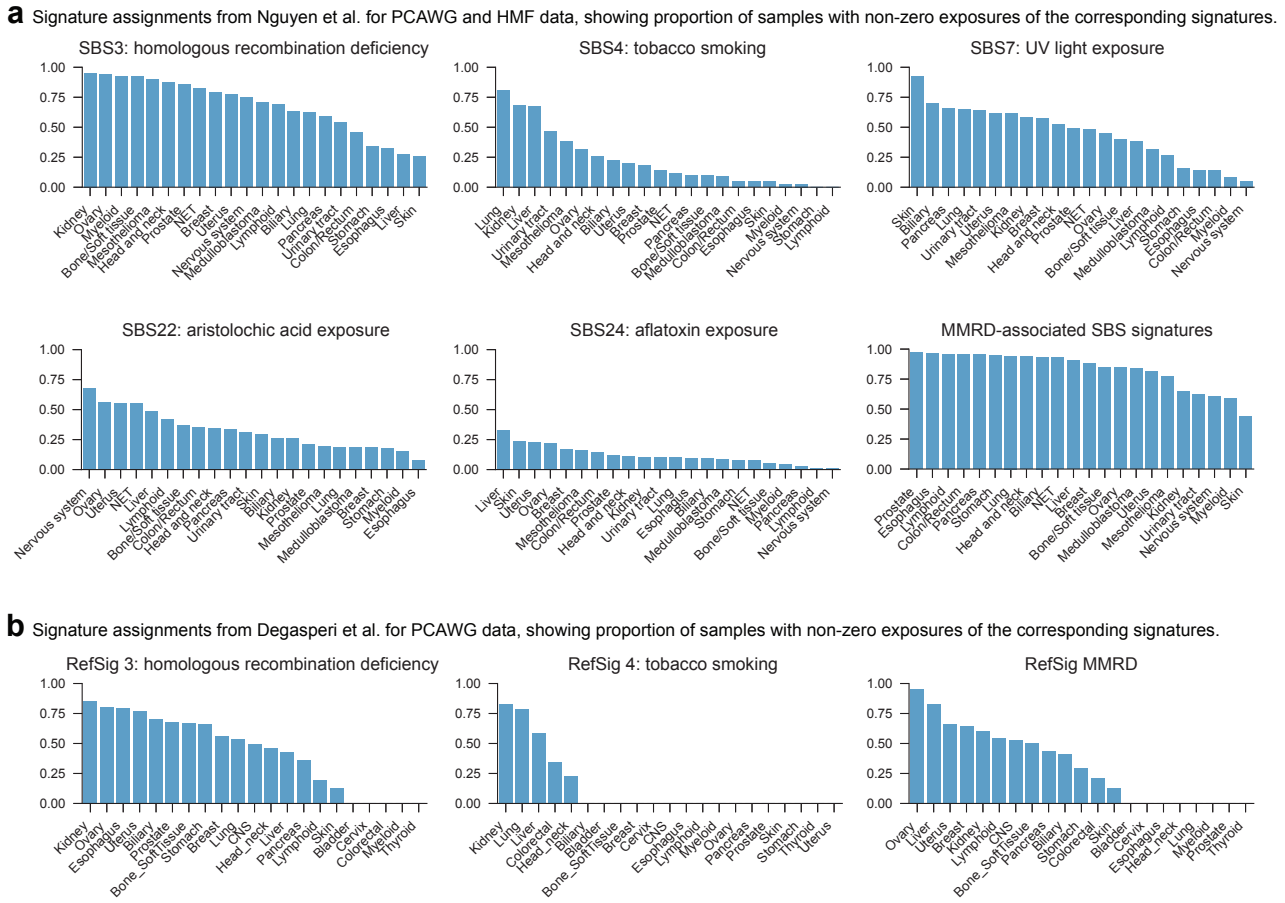
[67] Steele, C. D. et al. Signatures of copy number alterations in human cancer. bioRxiv (2021) .

[68] Bao, L., Zhong, X., Yang, Y. & Yang, L. Mutational signatures of complex genomic rearrangements in human cancer. bioRxiv (2021) .

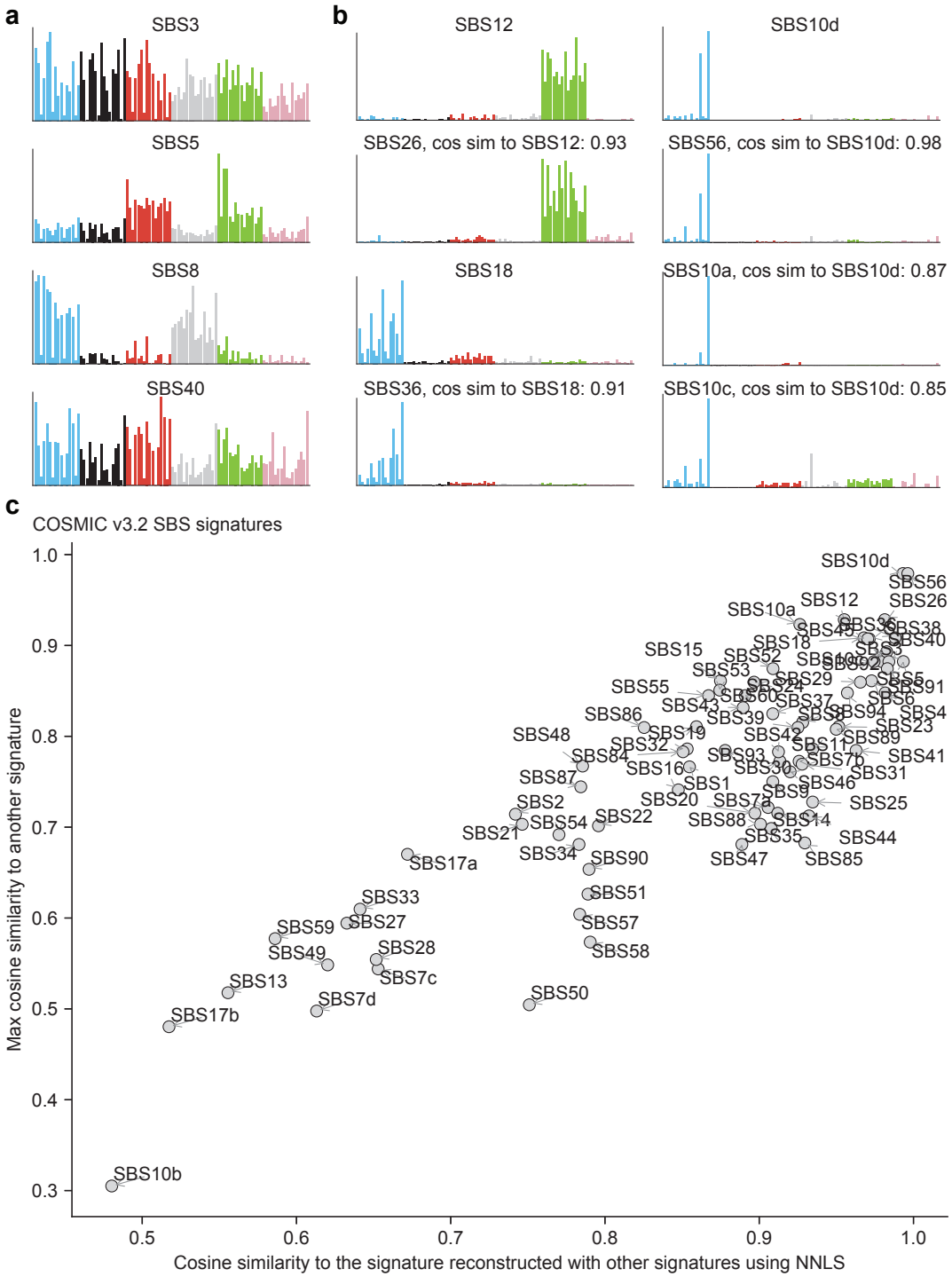
[69] Gao, G., Wan, W., Zhang, S., Redden, D. T. & Allison, D. B. Testing for differences in distribution tails to test for differences in 'maximum' lifespan. BMC Med Res Methodol **8**, 49 (2008) .

[70] Fujimoto, A. et al. Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types. Genome Res (2020) .

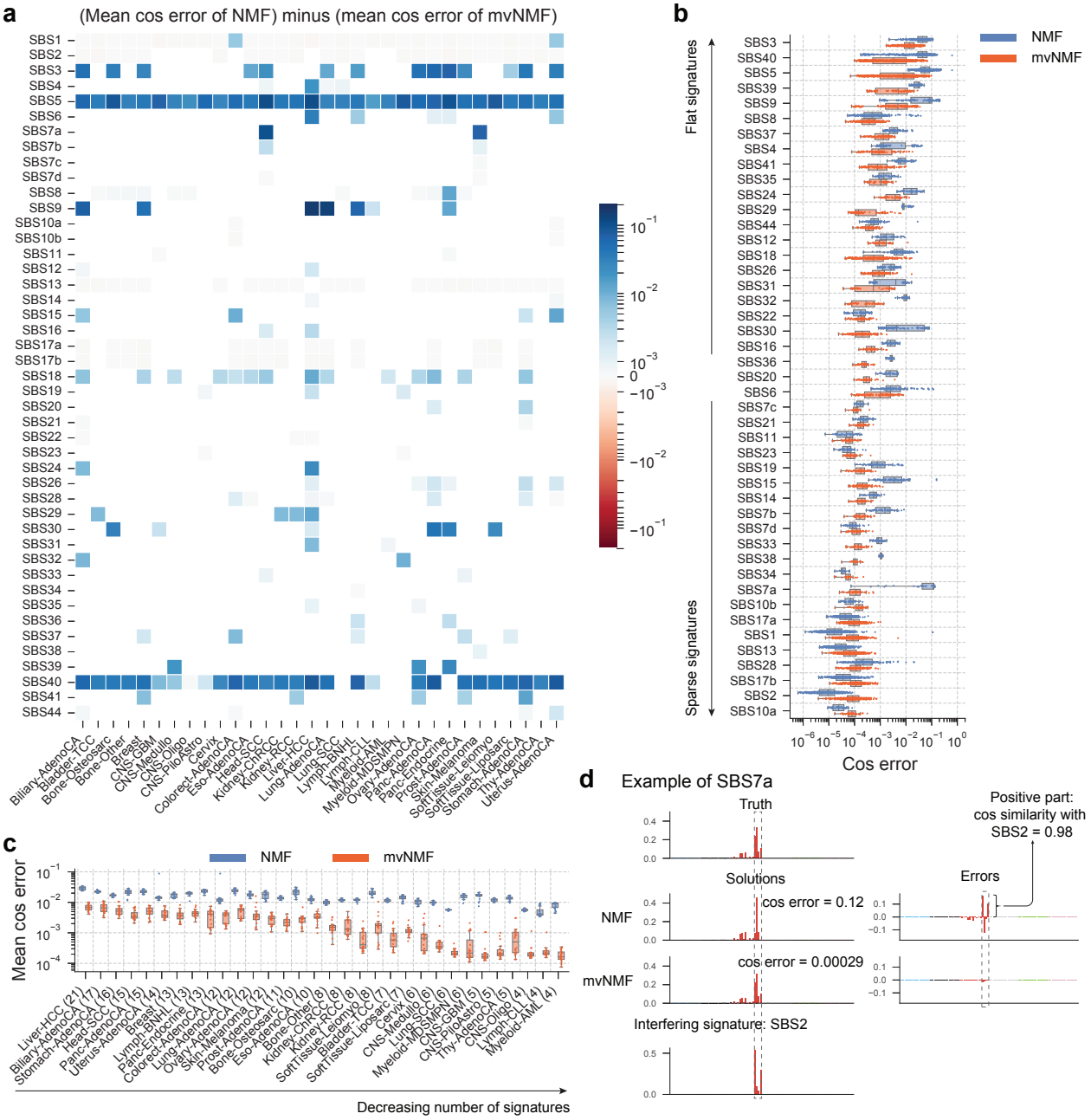
Appendix A Extended Data



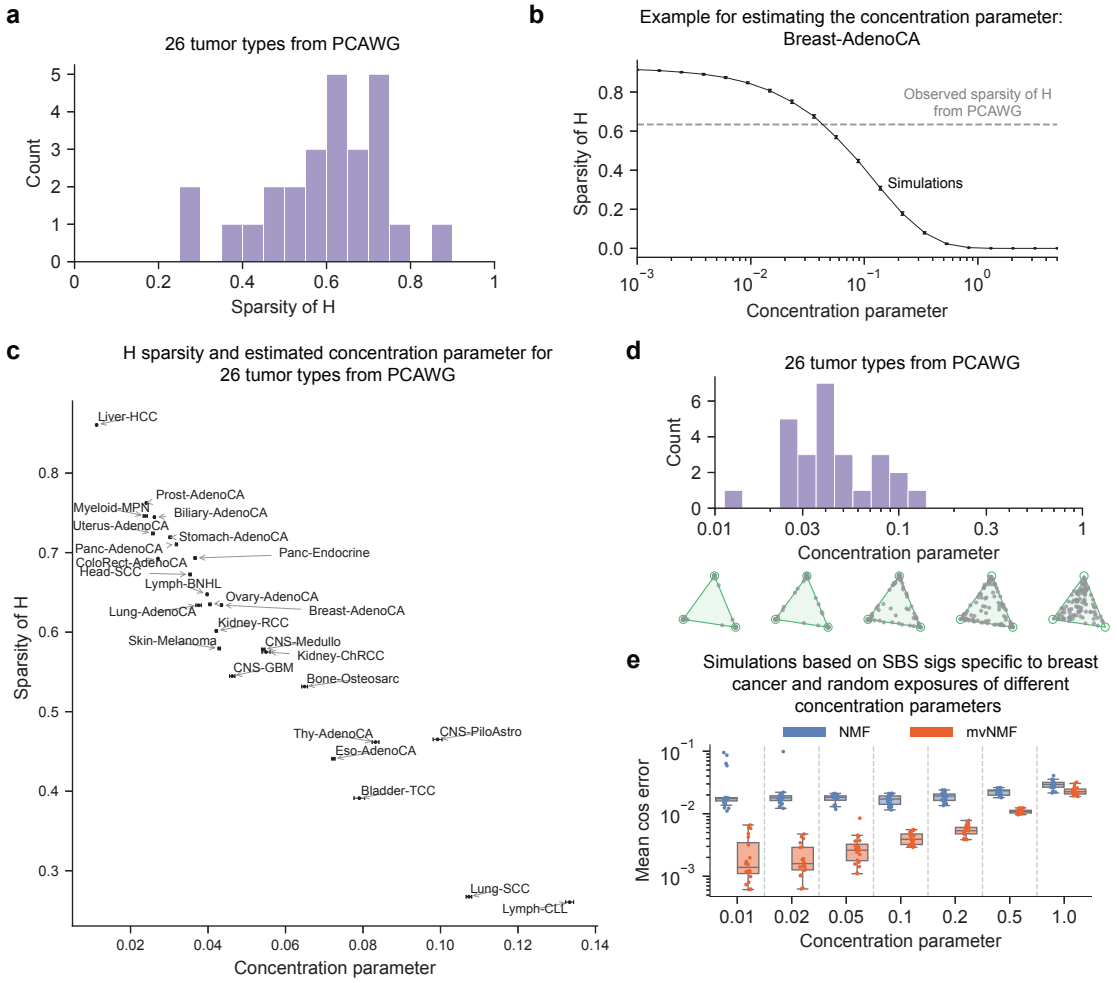
Extended Data Fig. 1 Examples of erroneous signature assignment in previous studies. **a.** Examples of signature assignment results obtained from Nguyen et al. [29], showing the proportion of samples with non-zero exposures of the corresponding signatures for each tumor type. Results for PCAWG data and the Hartwig Medical Foundation (HMF) cohort are combined. SBS3 corresponds to homologous recombination deficiency (HRD) but is assigned to a large fraction of samples in many tumor types not known to exhibit HRD, such as kidney, myeloid, etc. Similar problems are observed for other signatures – tobacco smoking-associated SBS4 should be specific to lung, liver, and head and neck cancers; UV light exposure-associated SBS7 should be exclusively observed in skin cancers; aristolochic acid-associated SBS22 and aflatoxin-associated SBS24 should be specific to liver, biliary, and kidney cancers; and MMRD-associated signatures (SBS6, 14, 15, 20, 21, and 26) should be assigned to MMRD samples, which are prevalent only in colorectal, esophagus, stomach, and uterus cancers. **b.** Examples of signature assignment results obtained from Degasperi et al. [30] for PCAWG data. Degasperi et al. use a different catalog of signatures with the prefix RefSig. RefSig 3 is similar to COSMIC SBS3 and corresponds to HRD. RefSig 4 is similar to COSMIC SBS4 and corresponds to tobacco smoking. Two RefSig signatures – RefSig MMR1 and RefSig MMR2 – correspond to MMRD and are combined as RefSig MMRD in the plot. Similar problematic signature assignments as in (a) are observed, e.g., substantial over-assignment of RefSig 3 in many tumor types, RefSig 4 in kidney and colorectal cancers, and RefSig MMRD in tumor types without prevalent MMRD.



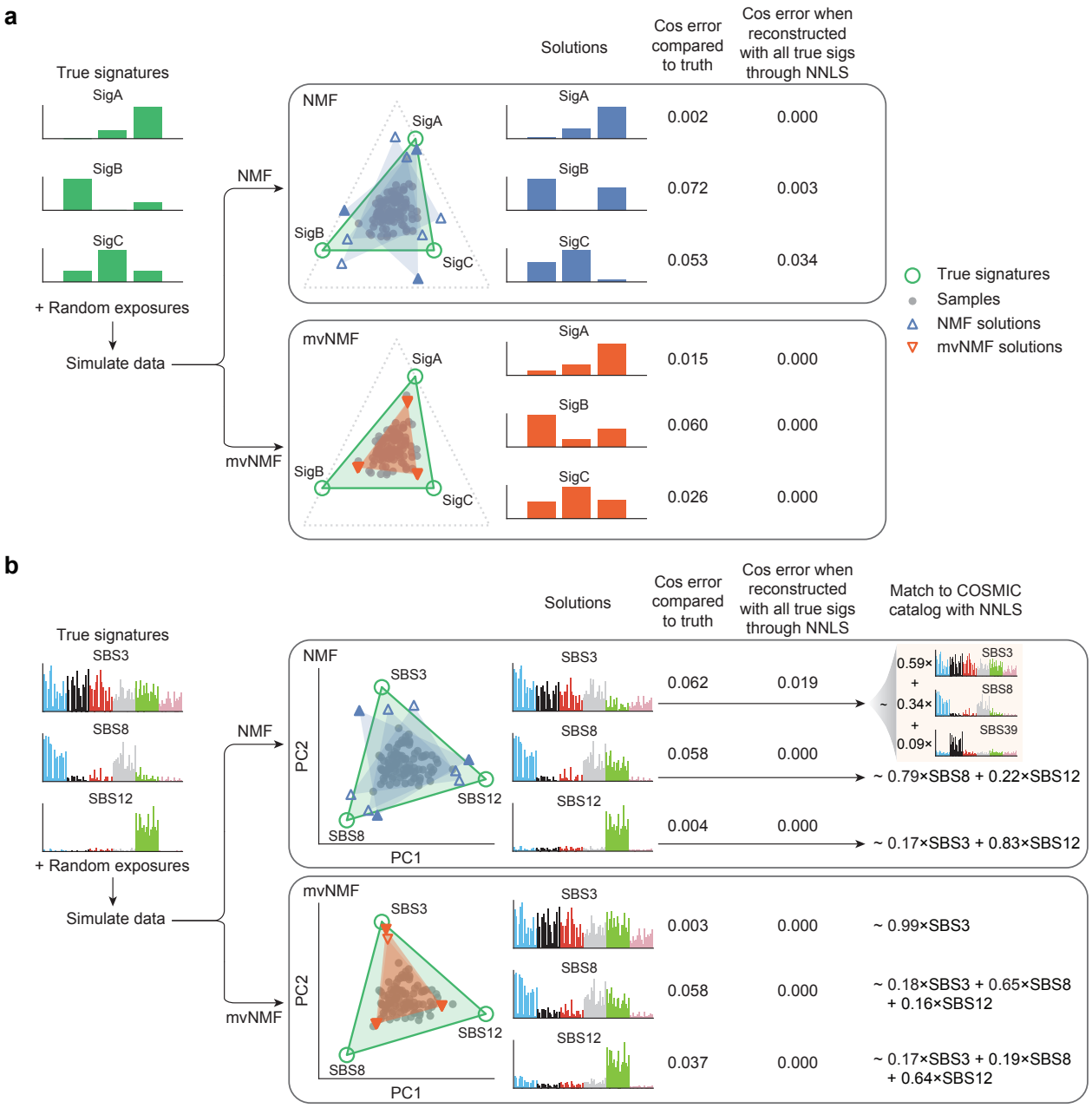
Extended Data Fig. 2 The COSMIC catalog contains signatures that are flat, overly similar, or can be expressed as linear combinations of each other. **a.** Examples of relatively flat signatures. **b.** Examples of similar signatures. SBS10d, 56, 10a, and 10c are all similar to each other. SBS12 is similar to SBS26. SBS18 is similar to SBS36. **c.** Similarity between different SBS signatures in the COSMIC catalog. For each signature, the maximum cosine similarity to another signature in the catalog is plotted on the y-axis, and the cosine similarity to the signature reconstructed by fitting with all other signatures in the catalog through NNLS is plotted on the x-axis.



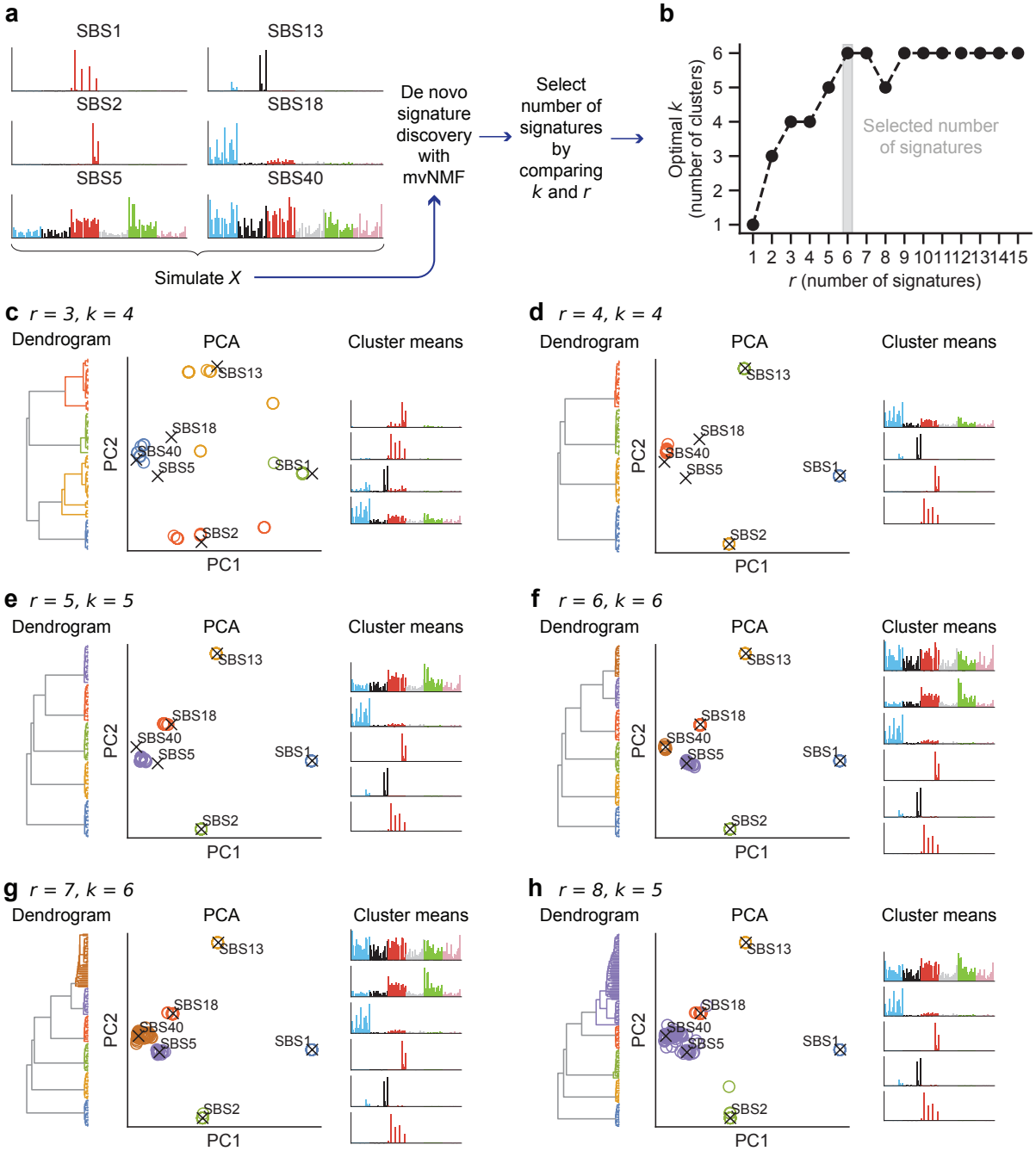
Extended Data Fig. 3 mvNMF improves the accuracy of *de novo* signature discovery in simulated datasets with SBS signatures. Synthetic datasets are simulated from tumor type-specific SBS signatures and Dirichlet-distributed exposures for 32 tumor types with 20 replicates. Each dataset contains 200 samples and on average 5,000 mutations per sample (Poisson-distributed). NMF and mvNMF are then applied for *de novo* signature discovery, assuming that the number of signatures is known. Finally, the discovered signatures are compared to the true ones, and their discrepancies are quantified by cosine errors. **a.** Heatmap of the difference between NMF- and mvNMF-derived cosine errors. Each element represents the mean of 20 independent simulations. **b.** NMF- and mvNMF-derived cosine errors for different SBS signatures, sorted by standard deviation of the corresponding signature spectrum. Data from different tumor types are collapsed. Same as Fig. 2c and included for completeness. **c.** NMF- and mvNMF-derived cosine errors for different tumor types, sorted by the number of signatures present in the corresponding tumor type. Data from different signatures are averaged within each tumor type. **d.** An example comparing the performance of NMF and mvNMF on identifying SBS7a. The NMF solution of SBS7a receives a large cosine error. The error spectrum indicates interference from SBS2 coexisting in the dataset. By comparison, mvNMF does not suffer from the SBS2 interference and is able to discover SBS7a accurately.



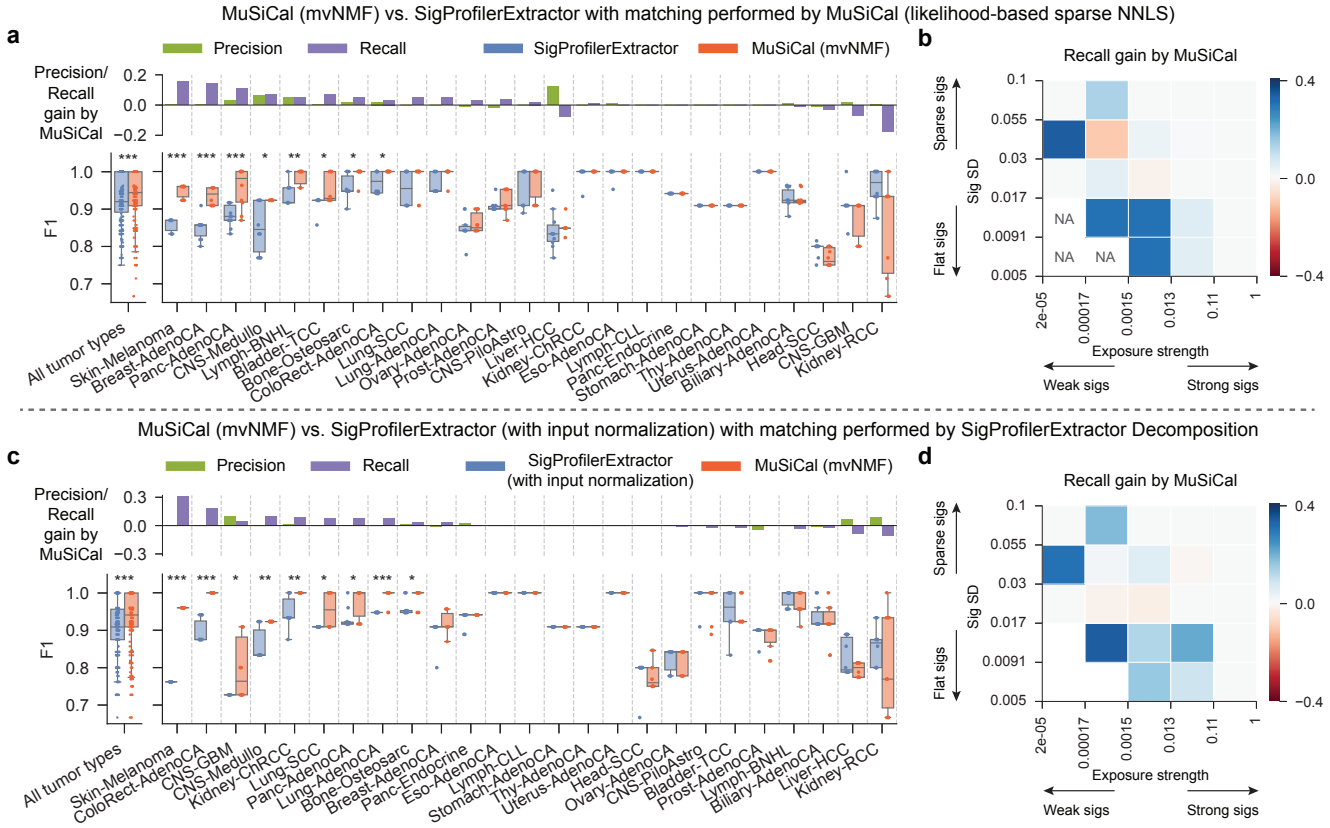
Extended Data Fig. 4 mvNMF is suitable for *de novo* signature discovery from cancer data. Theoretically, mvNMF is guaranteed to recover the true underlying signatures if the exposure matrix H satisfies the so-called sufficiently scattered condition [46, 50, 51]. Intuitively, the sufficiently scattered condition requires that only a few signatures are active in each sample [46, 50, 51]. Although this condition is seemingly reasonable for tumor samples, we still set out to investigate to what extent it is satisfied in real cancer data. Empirically, the sufficiently scattered condition is more likely to be satisfied when H is sparse enough [50, 51]. We therefore study the tumor type-specific exposure matrices from PCAWG results [14]. **a.** Distribution of sparseness (as quantified by the fraction of zero elements) for PCAWG exposure matrices of 26 tumor types with at least 20 samples. The exposure matrices of PCAWG tumors contain 53% of zero elements on average (range 22-86%), indicating strong sparsity. **b.** An example demonstrating how the concentration parameter can be estimated from exposure sparseness. Assuming that the sample exposures in H follow a symmetric Dirichlet distribution, the concentration parameter α, defined as the value of the single Dirichlet parameter used in all dimensions, is directly related to the sparseness of H. A concentration parameter of 1 results in exposure vectors being uniformly distributed over the simplex, and thus non-sparse, whereas a concentration parameter much smaller than 1 corresponds to sparse exposures. Therefore, in order to obtain a more quantitative insight, we estimate α from exposure sparseness for PCAWG tumors. To illustrate how the estimation is performed, we take the exposure matrix of Breast-AdenoCA from PCAWG results as an example. The exposure matrix has a sparsity level of 0.63 (gray dashed line), as quantified by the fraction of zero elements. Synthetic exposure matrices with the same number of samples and mutation counts are simulated from Dirichlet distributions with different concentration parameters, and the resulting exposure sparseness is plotted (black line, error bar indicating standard deviation of 20 independent simulations). A monotonic and tight relation between exposure sparseness and the concentration parameter is observed, allowing the concentration parameter to be determined using Brent's root-finding algorithm. Because of the randomness from simulations, the solver is run 10 times independently, and the mean solution is taken. For Breast-AdenoCA, α is estimated to be 0.043, with a standard deviation of 0.0002. **c.** Exposure sparseness and estimated concentration parameters for 26 PCAWG tumor types with at least 20 samples. Error bars indicate the standard deviation of estimated concentration parameters from 10 independent runs. **d.** Distribution of estimated concentration parameters for PCAWG exposure matrices of 26 tumor types with at least 20 samples. Low-dimensional examples are also shown to illustrate the sparsity level at different concentration parameters of 0.01, 0.03, 0.1, 0.3, and 1, respectively. The estimated α is 0.05 on average (range 0.01-0.13) for PCAWG exposure matrices – much smaller than 1 – again indicating strong sparsity. Therefore, somatic mutations from cancer data are highly likely to satisfy the sufficiently scattered condition, justifying the application of mvNMF. **e.** mvNMF improves the accuracy of *de novo* signature discovery for a wide range of exposure sparseness relevant for cancer data. Simulations in Fig. 2c and Extended Data Fig. 3 are repeated for different α values between 0.01 and 1. An example is shown for simulations based on SBS signatures specifically present in breast cancer. Cosine errors averaged across these signatures are plotted for different values of concentration parameters.



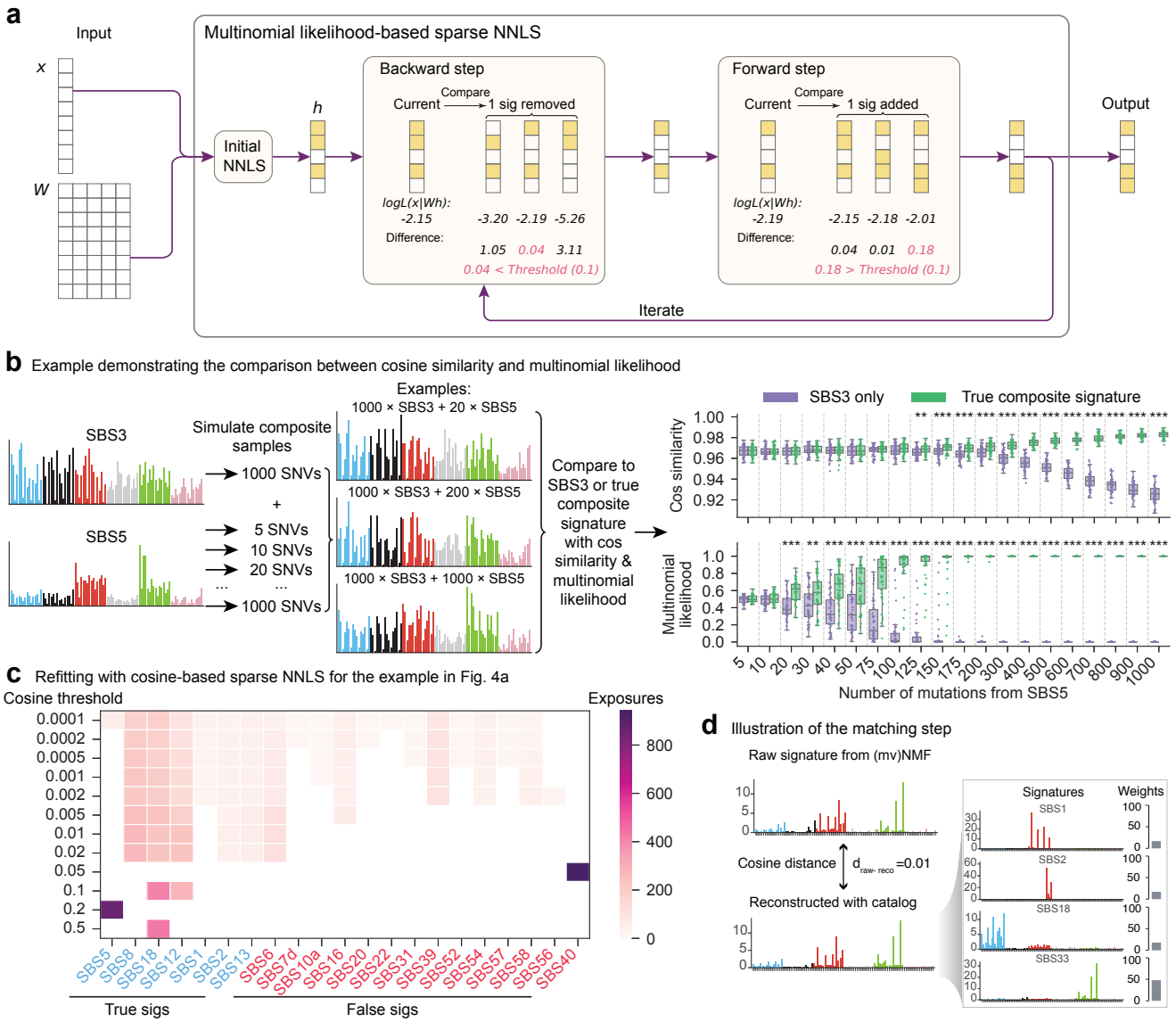
Extended Data Fig. 5 mvNMF is favored over NMF even when the exposure matrix is non-sparse, as mvNMF always ensures that the solutions are non-negative mixtures of true signatures. a. An example with three-dimensional signatures. Synthetic samples are simulated as in Fig. 2a, except that a large concentration parameter $\alpha = 5$ is used for Dirichlet-distributed random exposures. NMF and mvNMF are then applied to recover the signatures with three different initializations (shared between NMF and mvNMF). The resulting signature solutions are visualized on the simplex $x + y + z = 1$ as scatter plots. The spectra of example solutions (highlighted by filled triangles for NMF and mvNMF solutions) are shown. In this example, the exposure matrix is dense enough such that all signatures are active in each single sample. As a result, the signatures discovered by mvNMF can be inaccurate if directly compared to the underlying truth due to shrinkage from the volume regularization (see the cosine errors between the solutions and the corresponding true signatures annotated next to the example spectra). However, the mvNMF solutions are guaranteed to be non-negative mixtures of the true signatures, as indicated by the zero cosine errors between the solutions and the signatures reconstructed with all three true signatures through NNLS. By contrast, NMF solutions have no such guarantees: they can be arbitrarily distorted and thus cannot be reconstructed with the true signatures through NNLS. **b.** An example with real SBS signatures. The simulations are performed as in (a), except that real SBS signatures (SBS3, 8, and 12) are used, and the signature solutions are visualized in the PCA space as scatter plots. Again, mvNMF solutions are ensured to be non-negative mixtures of true signatures, whereas NMF solutions are not. Consequently, when matched to the entire COSMIC catalog through NNLS (with an exposure cutoff of 0.05), mvNMF solutions are correctly identified as combinations of true signatures. By contrast, an additional false signature SBS39 is identified for the NMF solution of SBS3 to compensate for the negative algorithmic distortions.



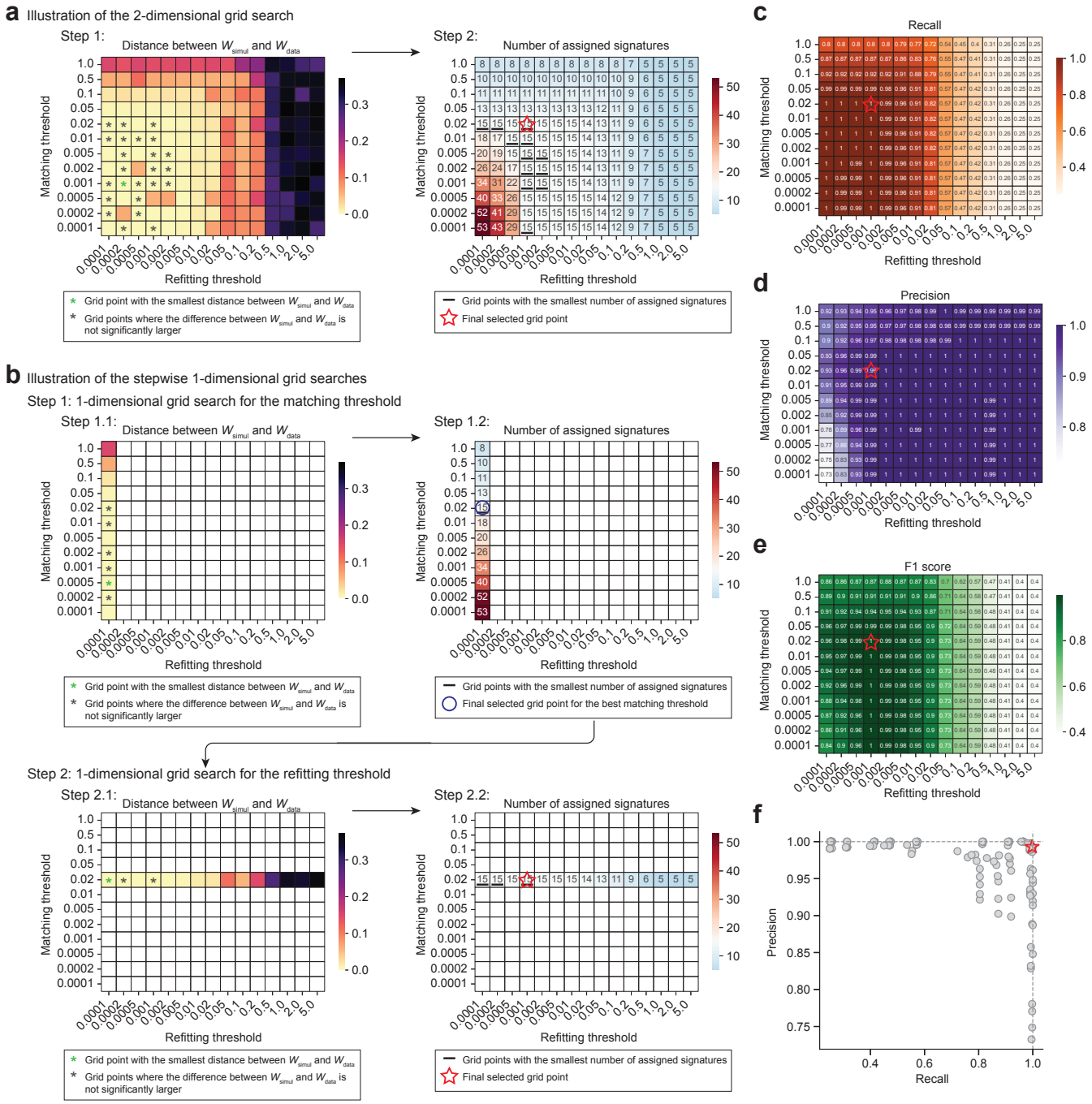
Extended Data Fig. 6 An example illustrating the method for selecting the number of signatures r. a. Data simulation scheme. A synthetic dataset is simulated from Dirichlet-distributed exposures and 6 SBS signatures (SBS1, 2, 5, 13, 18, and 40) believed to be present in cervical cancers according to PCAWG results. The dataset contains 200 samples and on average 5000 mutations per sample. mvNMF is then run for $r = 1$ to 15 separately, with 20 independent replicates from different initializations for each r value. **b.** Selecting the number of signatures by comparing k and r. *De novo* signatures from solutions with the same r are clustered into k clusters, where k is determined by the gap statistic [54, 55]. The greatest r such that $k = r$ is chosen as the final number of signatures. In this example, the correct number of signatures ($r = 6$) is selected. **c-h.** Details of mvNMF solutions at different r values. For each r value, the mvNMF-derived *de novo* signatures from 20 independent runs are clustered and visualized as dendrograms and PCA plots. The signatures corresponding to cluster means are also shown. When r is smaller than the true number of signatures (e.g., $r = 3$), the mvNMF solutions can be unstable, resulting in $k > r$. When r is greater than the true number of signatures (e.g., $r = 7$), mvNMF may produce redundant signatures, resulting in $k < r$. Together, these observations suggest that the greatest r with $k = r$ is a reasonable estimate of the true number of signatures.



Extended Data Fig. 7 MuSiCal outperforms the state-of-the-art algorithm SigProfilerExtractor for *de novo* signature discovery. a, b. Same as Fig. 3b, c, but the matching algorithm from MuSiCal is used instead of the Decomposition module from SigProfilerExtractor. MuSiCal utilizes the likelihood-based sparse NNLS for matching. In order to minimize the impact of the matching step on evaluating *de novo* signatures, the likelihood threshold is separately adjusted for each solution to achieve the optimal F1 score, for both MuSiCal and SigProfilerExtractor. **c, d.** Same as Fig. 3b, c, but the input normalization from SigProfilerExtractor is turned on. Note that the preprocessing module from MuSiCal is still skipped in the comparison. Therefore, even with input normalization, SigProfilerExtractor still performs worse than MuSiCal without preprocessing.

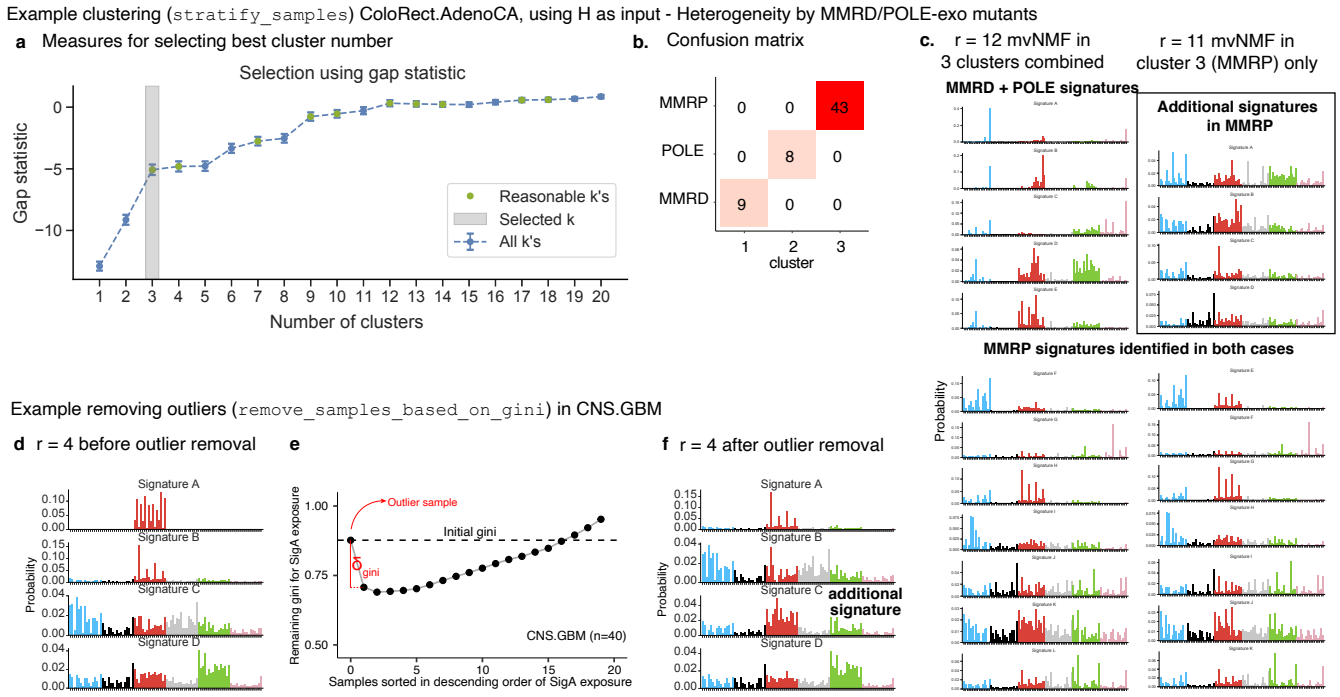


Extended Data Fig. 8 Matching and refitting with multinomial likelihood-based sparse NNLS. a. Diagram of the likelihood-based sparse NNLS algorithm. See Methods for details. **b.** An example demonstrating that multinomial likelihood is more powerful than cosine similarity for separating similar signatures. SBS3 and SBS5 are used to simulate synthetic samples. All samples contain 1000 SNVs contributed by SBS3 as well as varying numbers of SNVs from SBS5. Multinomial likelihood and cosine similarity are then applied to distinguish whether the sample spectra are generated from the correct (SBS3 + SBS5 with appropriate weights) or incorrect (pure SBS3) underlying signatures. The problem is expected to be difficult with few SNVs from SBS5, as the sample spectra will be dominated by SBS3. Indeed, cosine similarity fails to separate the two underlying signatures when there are less than 100 SNVs from SBS5, corresponding to an SBS5 exposure of $100/(100+1000) = 9\%$. By comparison, multinomial likelihood achieves statistically significant separation down to 20 SNVs from SBS5, corresponding to an SBS5 exposure of 2%. *: $p < 0.05$. **: $p < 0.005$. ***: $p < 0.0005$. p-values are calculated with t-tests. **c.** Same as Fig. 4b, but refitting is performed using cosine similarity combined with the same bidirectional stepwise algorithm as in (a). Notably, there is no threshold with which the set of active signatures are identified correctly. Also, solutions with cosine similarity do not possess the desired property of continuity as in Fig. 4b. For example, even when the threshold is overly small, true signatures (e.g., SBS5) can be missed, while when the threshold is overly large, false signatures (e.g., SBS40) can be discovered instead of the strongest true signatures. **d.** Illustration of the matching step. MuSiCal uses the same likelihood-based sparse NNLS for the matching step, where a *de novo* signature is decomposed as a non-negative mixture of known signatures in the catalog.

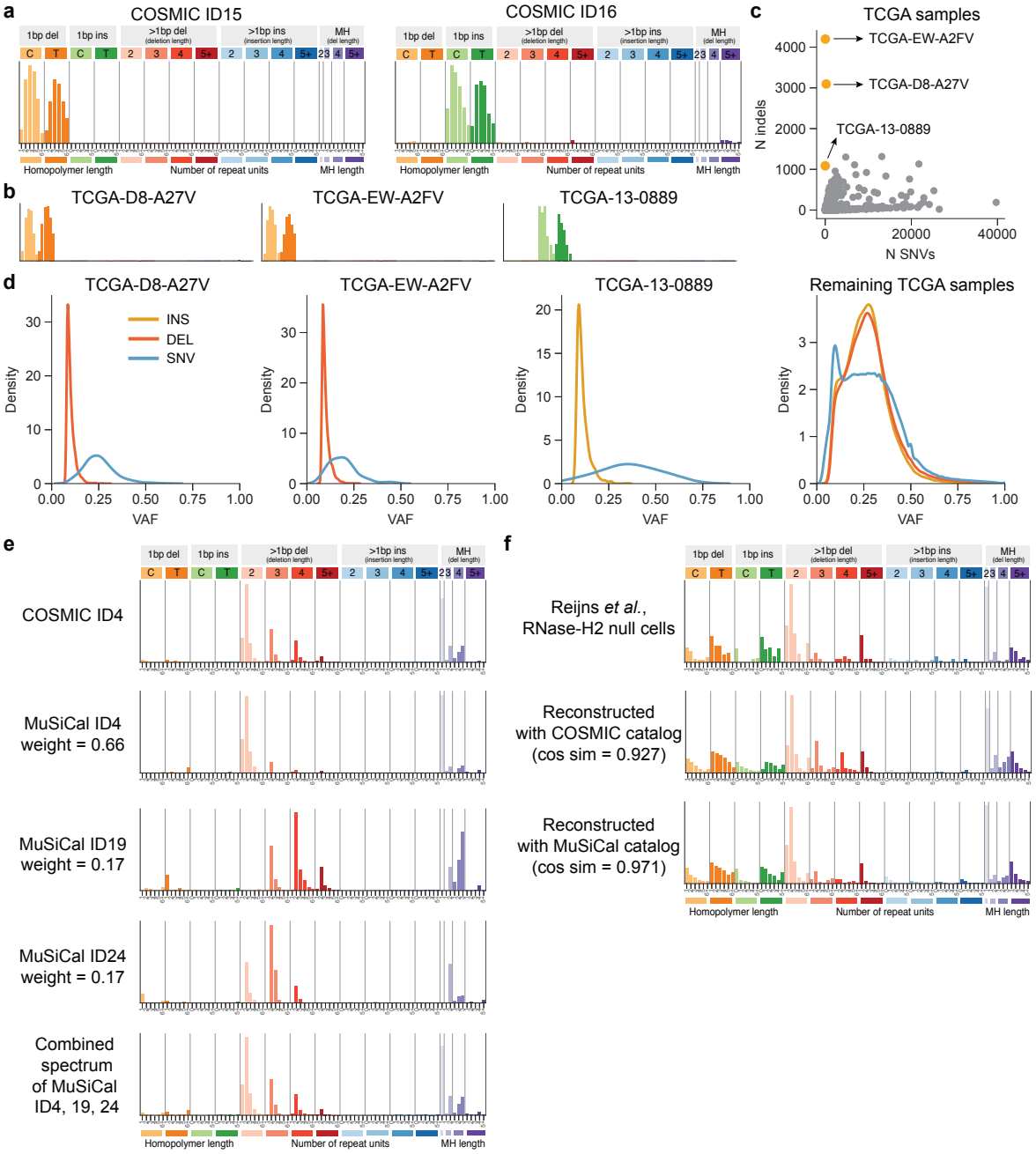


Extended Data Fig. 9 (Continued on the following page.)

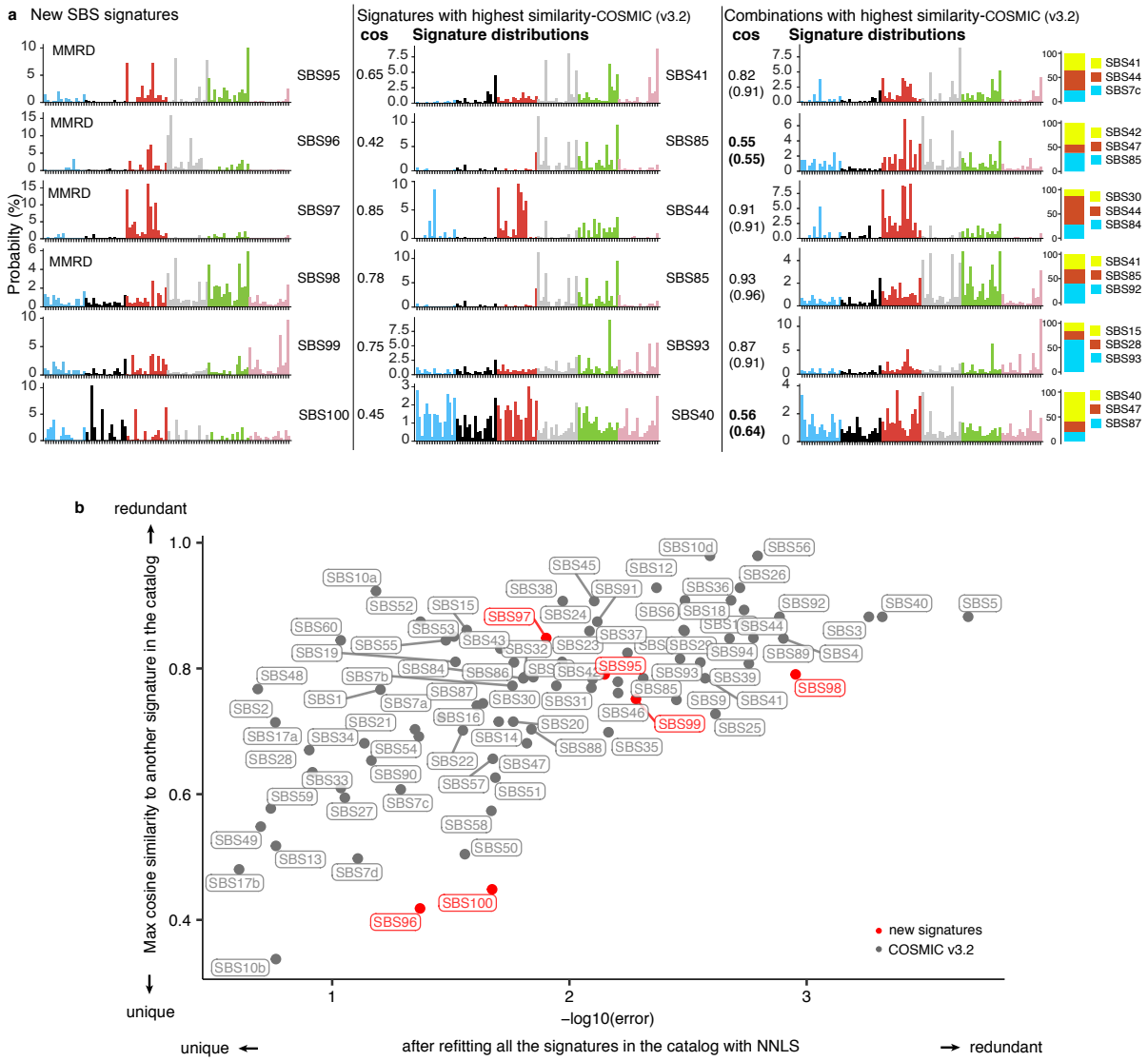
Extended Data Fig. 9 An example based on the PCAWG Skin Melanoma dataset to illustrate post-discovery parameter optimization through data-driven simulations. The *in silico* validation/optimization module can be used to select the optimal post-discovery parameters by comparing the original data with data-driven simulations. To demonstrate this approach, a realistic synthetic dataset based on PCAWG skin melanomas is used, such that the true underlying signatures (15 signatures) and exposures are known to evaluate the results. An initial *de novo* signature discovery is first performed to derive W_{data} and H_{data} (see main text and Fig. 5a for explanation of the notations). Then, the two likelihood thresholds used in matching and refitting, respectively, are optimized through a two-dimensional grid search with data-driven simulations. Specifically, for each grid point (i.e., pair of likelihood thresholds), X_{simul} is simulated from the corresponding signature assignments W_{s} and H_{s} and then subject to *de novo* signature discovery again, resulting in W_{simul} and H_{simul}. The optimal grid point is selected based on the comparison between W_{simul} and W_{data}. Alternatively, to reduce computation time, two separate one-dimensional grid searches can be performed for the matching and refitting thresholds, respectively. **a.** Illustration of the two-dimensional grid search for jointly optimizing matching and refitting thresholds. In the first step, we inspect the mean cosine distance between W_{simul} and W_{data} (left panel). The grid point with the smallest distance (marked by green '*') is selected as a candidate solution. Additional grid points (marked by black '*') are further included in the candidate list if the corresponding difference between W_{simul} and W_{data} is not significantly larger (see Methods for details). In the second step, we inspect the number of assigned signatures (right panel). Among the candidate grid points selected in the first step, the one with the smallest number of assigned signatures is selected. In this case, there are multiple such grid points (marked by '−'). We select the one with the largest thresholds, i.e., the sparsest signature assignment (marked by a red star). **b.** Illustration of the alternative approach using stepwise one-dimensional grid searches. In the first grid search (top two panels), we optimize the matching threshold while fixing the refitting threshold to a small value (0.0001). In the second grid search (bottom two panels), we further optimize the refitting threshold while fixing the matching threshold to its optimal value (0.02). The best grid point is selected with the same procedure described in (a) for both grid searches. In this example, the joint approach (a) and the stepwise approach (b) select the same optimal grid point. **c-e.** Recall (c), precision (d), and F1 score (e) of signature assignments at different grid points. Recall and precision are calculated by comparing nonzero or zero entries in H_{s} with those in the true exposure matrix. The selected optimal grid point in (a) and (b) maximizes the F1 score, justifying the approach of parameter optimization by comparing W_{simul} and W_{data}. Note that different pairs of likelihood thresholds for matching and refitting can lead to the same signature assignment results. Therefore, multiple grid points achieve the best F1 score. **f.** Precision vs. recall of signature assignments at different grid points.



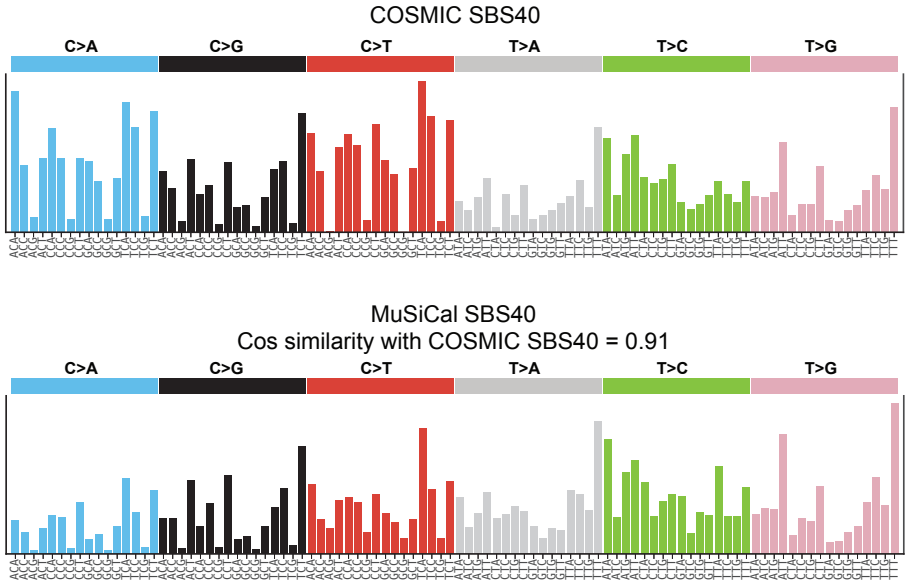
Extended Data Fig. 10 Examples illustrating cohort stratification and outlier removal with the preprocessing module. **a-c.** An example with the PCAWG ColoRect.AdenoCA dataset to demonstrate that cohort stratification can benefit *de novo* signature discovery. After an initial signature discovery with the entire dataset of 60 samples, hierarchical clustering is performed using the *de novo* exposure matrix H, and 3 clusters are selected by the gap statistic, as shown in panel (a). The 3 clusters correspond to samples with MMRP, MMRD, and POLE mutations, respectively, as shown in panel (b). In panel (c), *de novo* signatures discovered from the entire dataset of 60 samples and the MMRP cluster (48 samples) are plotted, demonstrating improved sensitivity of signature discovery after cohort stratification. Specifically, 12 signatures are discovered when *de novo* discovery is performed with the entire dataset of 60 samples, while 11 are discovered with the MMRP cluster alone. Out of the 11 MMRP-specific signatures, only 7 are discovered before cohort stratification. Therefore, a separate run of *de novo* discovery with the MMRP cluster allows 4 more MMRP-specific signatures to be discovered. **d-f.** An example with the PCAWG CNS.GBM dataset to demonstrate that outlier removal can benefit *de novo* signature discovery. When signature discovery is performed with the entire dataset, 4 *de novo* signatures are discovered, as shown in panel (d), where Signature A corresponds to SBS11 associated with temozolomide treatment. When the *de novo* exposures are inspected with the Gini coefficient, a single outlier sample with an exceptionally strong exposure of Signature A is detected, as shown in panel (e). After removing this outlier and rerunning signature discovery with the remaining samples, an additional signature is discovered, as shown in panel (f), demonstrating improved sensitivity.



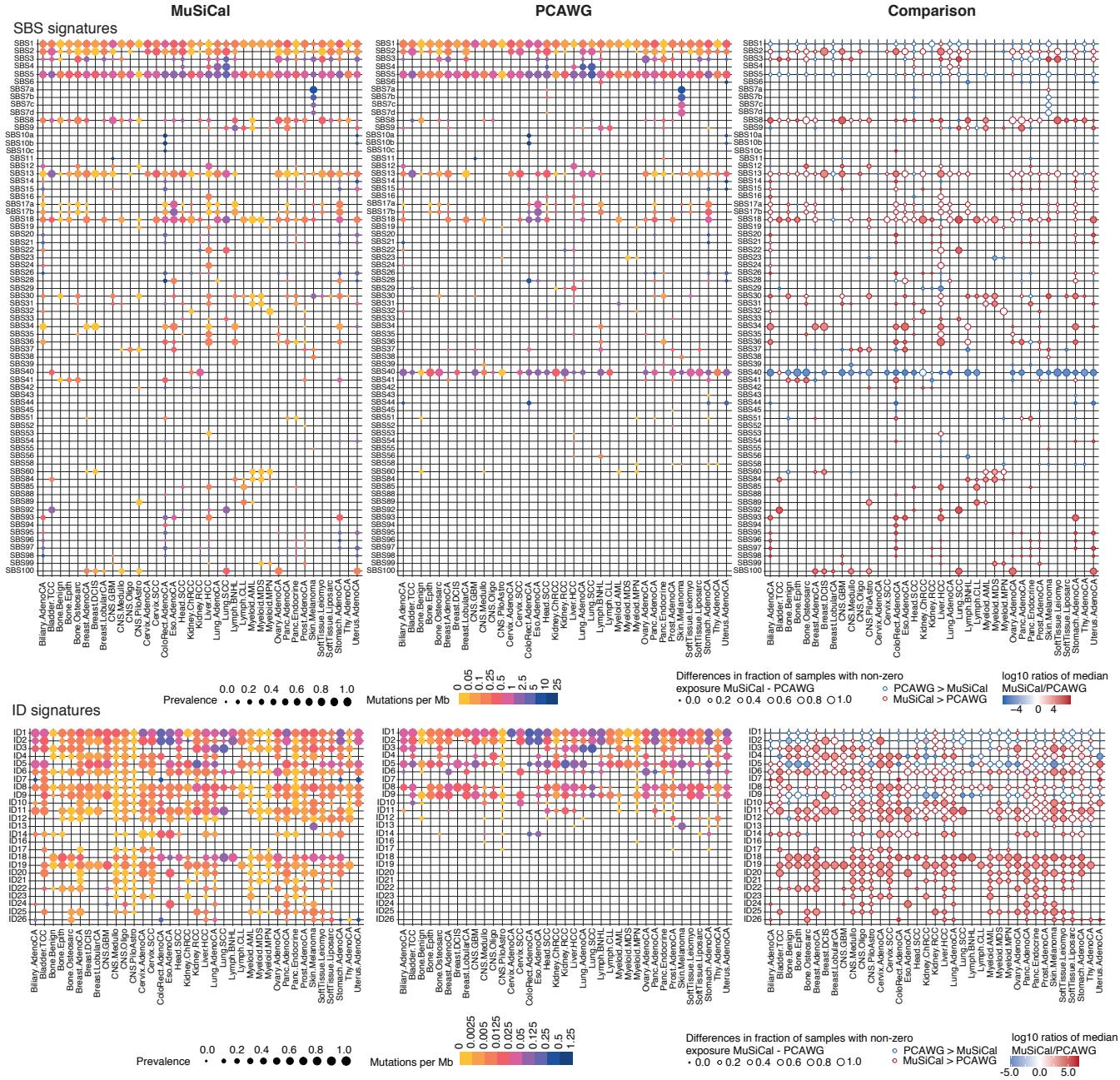
Extended Data Fig. 11 Further details related to the updated catalog of ID signatures. a. Spectra of COSMIC ID15 and 16. These two signatures are not discovered from our PCAWG reanalysis with MuSiCal. **b.** Indel spectra of the three TCGA whole-exome sequenced samples from which COSMIC ID15 and 16 are discovered by the PCAWG consortium [14]. **c.** Number of indels vs. number of SBSs for all TCGA samples, highlighting that these three samples have exceptionally high indel counts but low SBS counts. **d.** VAF distributions of indels and SBSs for these three samples are compared to those for the other TCGA samples. Indels in these three samples have particularly low VAF, suggesting that they are likely artifactual. **e.** COSMIC ID4 is resolved into multiple signatures in our updated catalog. COSMIC ID4 is decomposed using MuSiCal ID4, 19, and 24 with NNLS, and the corresponding exposures are annotated next to each of the MuSiCal signatures. The reconstructed signature has cosine similarity of 0.996 with COSMIC ID4 and is shown at the very bottom. **f.** The TOP1-associated ID spectrum observed in RNase-H2-null cells from [58] (top) is compared to the reconstructed spectra using the COSMIC (middle) and the MuSiCal catalog (bottom). The MuSiCal catalog better reconstructs the experimentally derived TOP1 signature. The TOP1 signature is more similar to MuSiCal ID4 (cosine similarity = 0.87) than COSMIC ID4 (cosine similarity = 0.83). Specifically, COSMIC ID4 contains longer (3- and 4-bp) deletions that are not observed in the TOP1 signature.



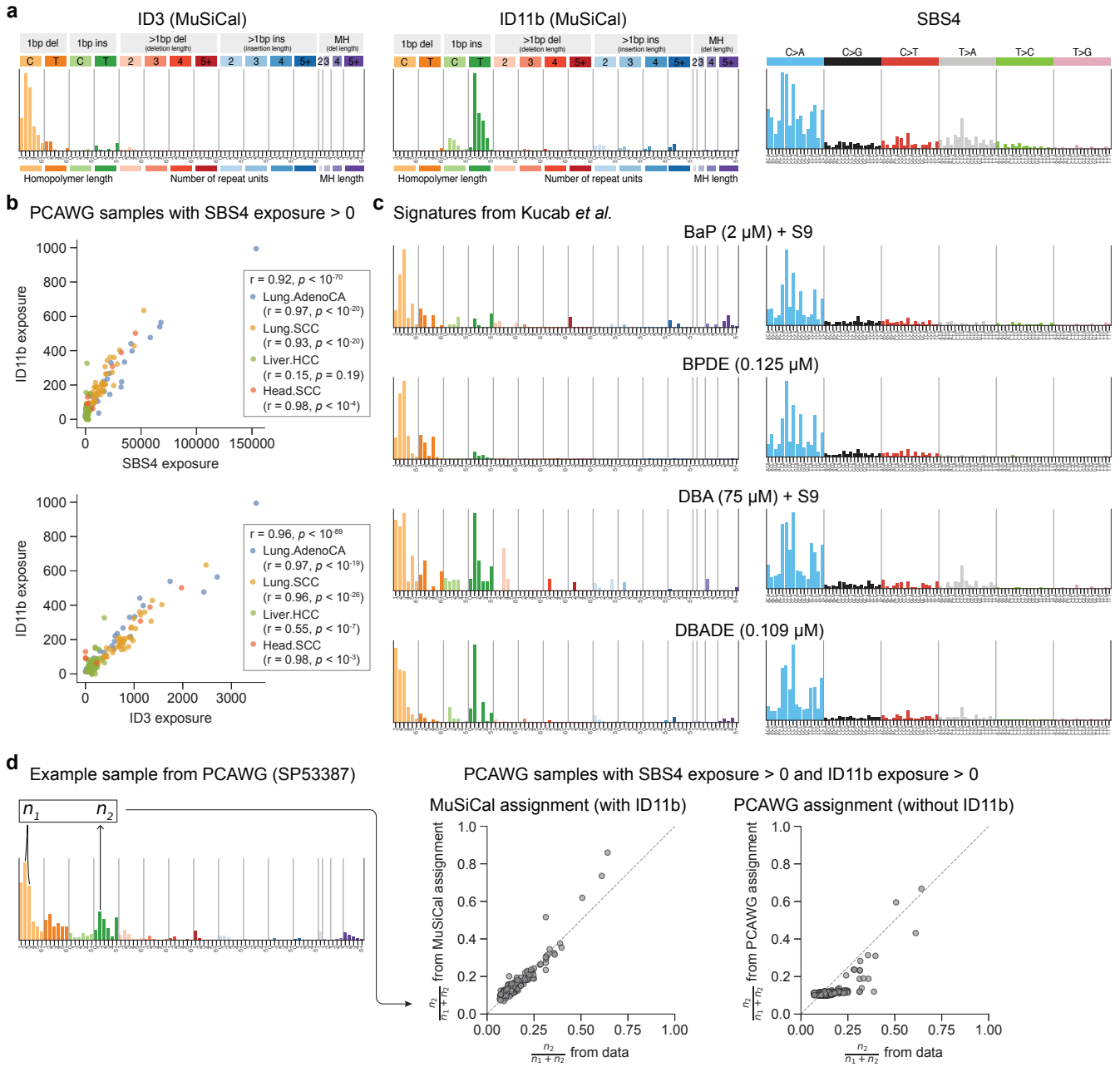
Extended Data Fig. 12 MuSiCal discovers 6 new SBS signatures from the PCAWG dataset. a. Spectra of the 6 new SBS signatures and comparison to the current COSMIC catalog. The 6 new SBS signatures discovered by MuSiCal are plotted to the left. In the middle, signatures from the COSMIC catalog with the highest cosine similarities to the 6 new ones, respectively, are plotted, and the corresponding cosine similarities are annotated next to the spectra. To the right, each of the 6 new SBS signatures is matched to the COSMIC catalog as a combination of at most 3 signatures through NNLS, and the reconstructed signature with the highest cosine similarity is plotted, with the corresponding cosine similarities annotated. The cosine similarities inside the parentheses are obtained from matching the new SBS signatures to the entire COSMIC catalog through NNLS without constraining to at most 3 signatures. Two new signatures, SBS96 and 100, are especially poorly reconstructed by COSMIC signatures and thus highlighted. **b.** Comparison of the 6 new signatures with those from the COSMIC catalog in terms of their uniqueness. To put the cosine similarities in (a) in context, two statistics are calculated for each signature in the COSMIC catalog as well as for the new ones – the maximum cosine similarity to another signature in the COSMIC catalog (y-axis), and the residual error after matching to all other signatures in the COSMIC catalog through NNLS (x-axis). Indeed, the 6 new signatures discovered by MuSiCal are not overly redundant compared to signatures already present in the COSMIC catalog. Two new signatures, SBS96 and 100, are especially unique compared to the majority of COSMIC signatures.



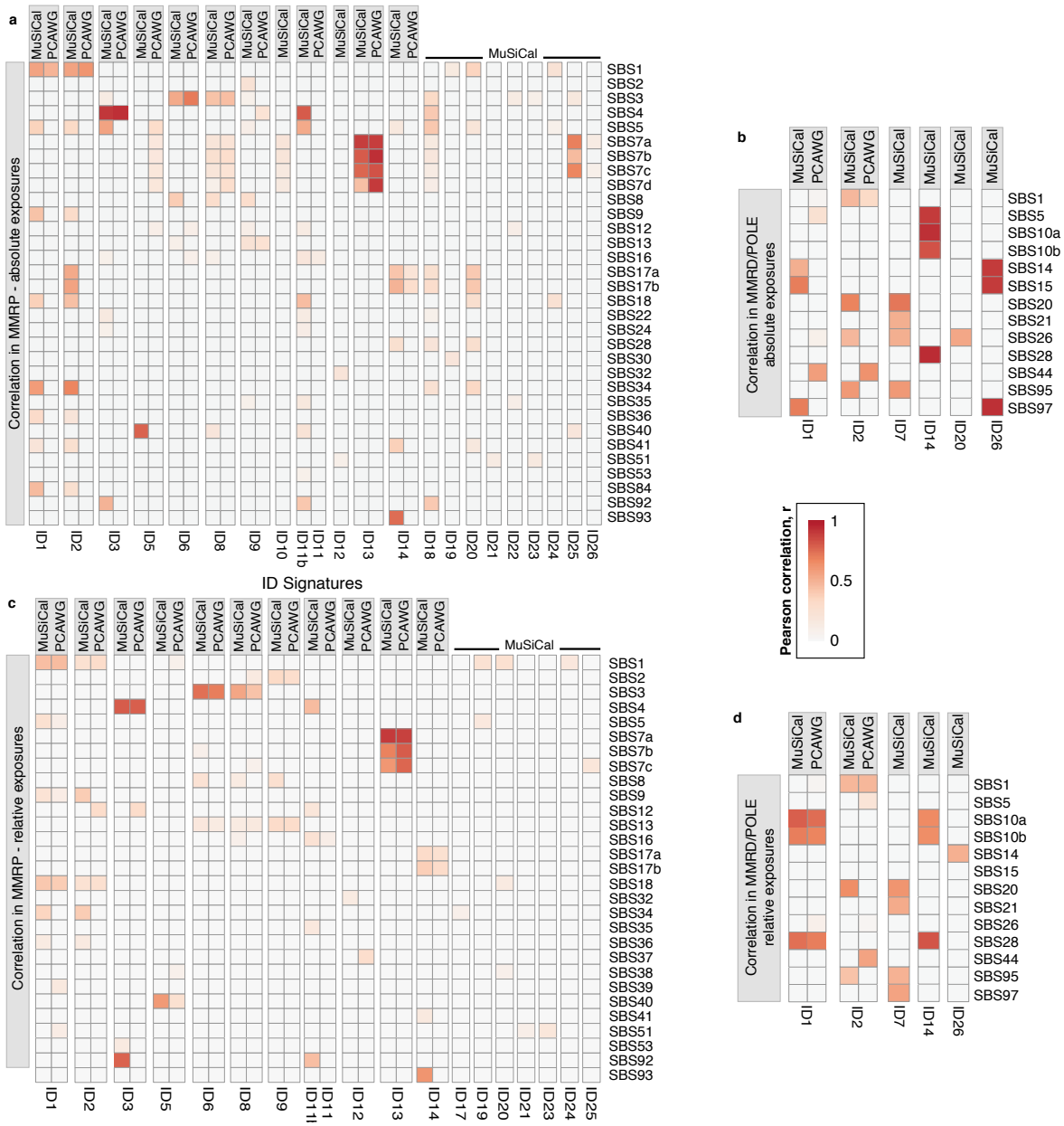
Extended Data Fig. 13 MuSiCal discovers a modified version of SBS40. SBS40 in the COSMIC catalog (top) is compared to the SBS40 discovered from our PCAWG reanalysis with MuSiCal (bottom). Compared to COSMIC SBS40, MuSiCal SBS40 has higher proportions of T>N mutations.



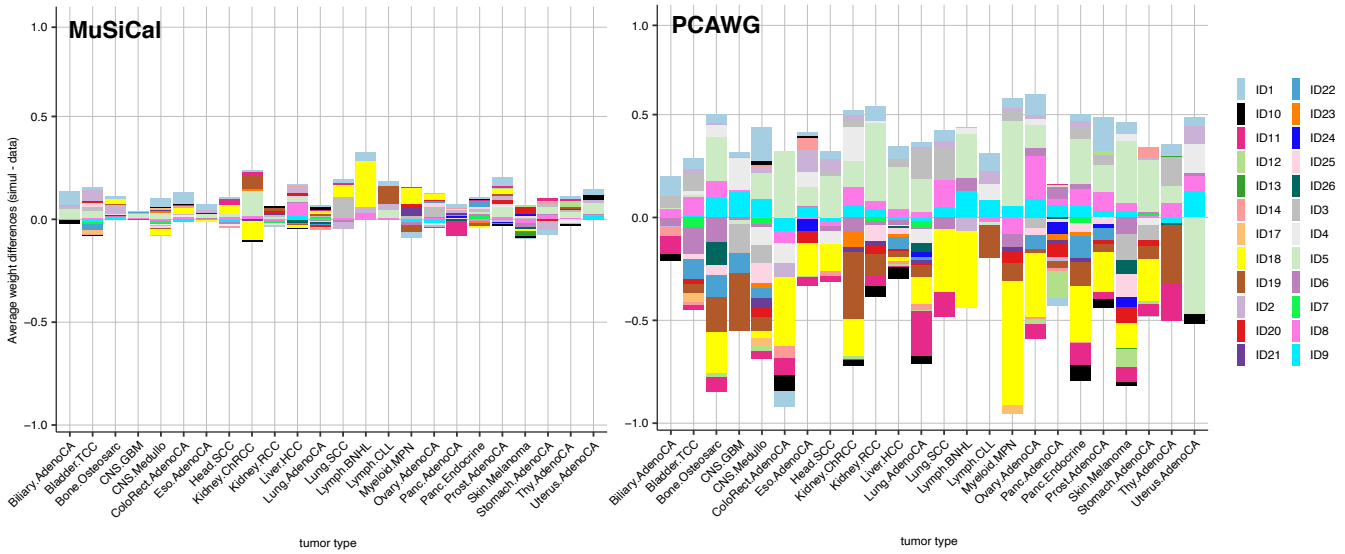
Extended Data Fig. 14 Comparison of signature assignments obtained by MuSiCal and the PCAWG consortium. Left panel is the same as Fig. 6b and included for visual comparison. Middle panel is obtained with exposure matrices produced by the PCAWG consortium [14], except that empty rows are added for the new signatures discovered by MuSiCal (SBS95-100 and ID19-27) as well as signatures recently added to the COSMIC catalog since the publication of [14] (SBS84-94). Note that exposures of ID11a and 11b are added together and presented as ID11 for MuSiCal results in order to facilitate comparison with PCAWG results. Right panel shows the comparison of MuSiCal and PCAWG assignments, where marker size represents the difference in proportion of samples with nonzero exposures of the corresponding signature for a given tumor type, and color indicates the log10 ratio of the median number of mutations per Mb contributed by the corresponding signature among samples with nonzero exposures.



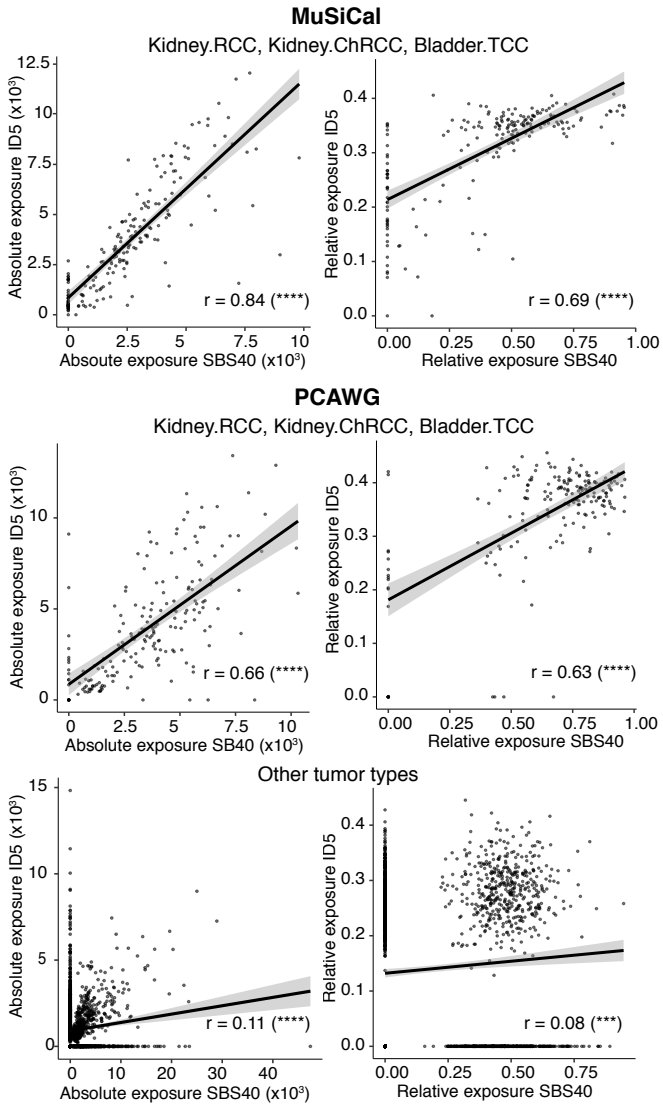
Extended Data Fig. 15 ID11b is potentially associated with tobacco smoking. a. Spectra of signatures known to be associated with tobacco smoking (ID3 and SBS4) and ID11b. **b.** ID11b exposure is correlated with both SBS4 and ID3 exposures. PCAWG samples related to tobacco smoking, as indicated by nonzero exposures of SBS4, are selected. The per-sample exposures are then plotted for ID11b vs. SBS4, and ID11b vs. ID3. **c.** ID (left) and SBS (right) signatures observed in cells exposed to different chemicals present in tobacco smoke. Mutation data is obtained from [16]. Background ID and SBS signatures averaged from control clones are removed. **d.** ID11b is required to explain the observed abundance of T insertions in PCAWG samples related to tobacco smoking. n_1 is defined as the number of C deletions at CC or CCC, characteristic of ID3, and n_2 is defined as the number of T insertions following a T, characteristic of ID11b (left). The ratio $\frac{n_2}{n_1+n_2}$ is then plotted for reconstructed spectra with MuSiCal assignment vs. observed spectra (middle), as well as for reconstructed spectra with PCAWG assignment vs. observed spectra (right). Without ID11b, the PCAWG assignment fails to explain the observed T insertions following a T.



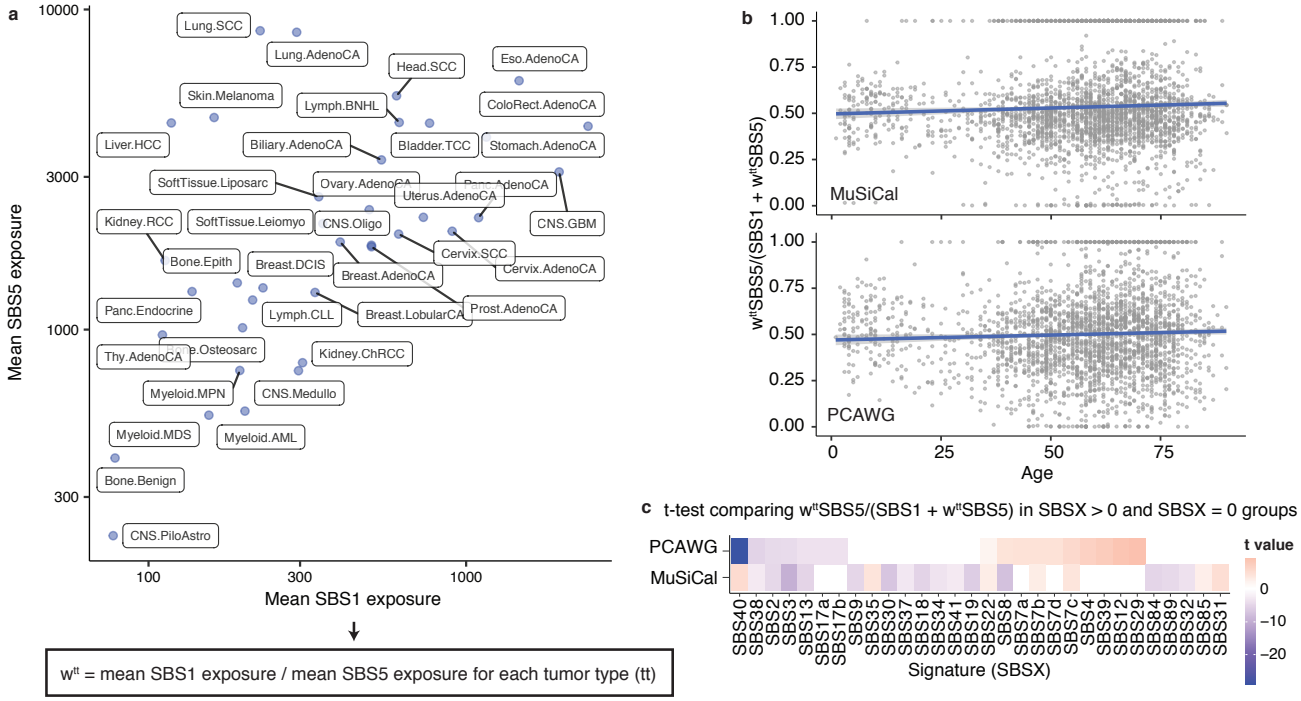
Extended Data Fig. 16 Comparison of SBS-ID associations in MuSiCal- and PCAWG-derived signature assignments. **a.** Same as Fig. 6c, except that correlations obtained using signature assignments produced by the PCAWG consortium are also shown for comparison for MMRP samples. **b.** Same as panel (a) but for MMRD, MMRD + POLE-exo mutant or POLE-exo mutant samples. **c-d.** Same as panels a-b, respectively, but for relative ID and SBS signature exposures.



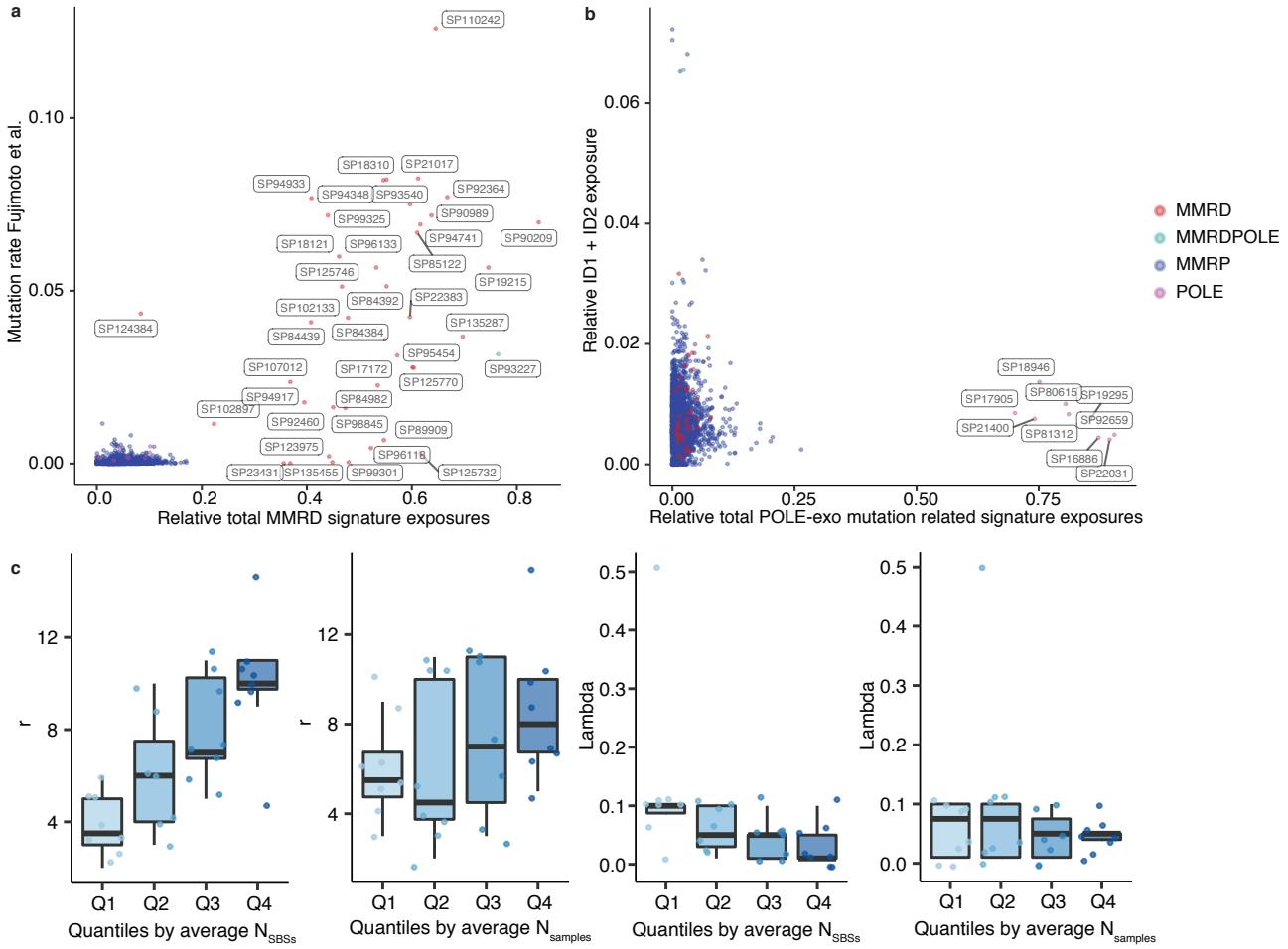
Extended Data Fig. 17 Comparison between MuSiCal- and PCAWG-derived ID signature assignments in terms of their consistency with data. Same as Fig. 7a, but for ID signatures.



Extended Data Fig. 18 Association between SBS40 and ID5. Exposures of SBS40 and ID5 are correlated for both MuSiCal- (top) and PCAWG-derived (middle) signature assignments in kidney and bladder cancers, which are the only tumor types where SBS40 is assigned by MuSiCal. PCAWG also assigns SBS40 in many other tumor types (bottom), where a strong correlation is not observed. Pearson correlation coefficients are shown for both absolute (left) and relative (right) exposures (normalized by mutation burden). ****: $p < 0.0001$. ***: $p < 0.001$.



Extended Data Fig. 19 Correlation between clock-like signatures, SBS1 and 5. a. Mean SBS5 exposure vs. mean SBS1 exposure in different tumor types, as obtained by MuSiCal, demonstrating the tumor type-specific accumulation rates of mutations from SBS1 and 5. To account for this tumor type-specificity, a normalization factor w^{tt} is calculated as the ratio between mean SBS1 exposure and mean SBS5 exposure for each tumor type (tt). The per-sample exposure of SBS5 is multiplied by w^{tt} from the corresponding tumor type when compared with that of SBS1. Exposures obtained by the PCAWG consortium are normalized in a similar way. **b.** Fraction of normalized SBS5 exposure relative to the sum of SBS1 and normalized SBS5 exposures vs. patient age. Since both SBS1 and SBS5 are clock-like and correlated, this fraction is expected to be a constant with respect to age and have a tight distribution (Fig. 7d). **c.** Comparison of samples with non-zero and zero exposures of different signatures suggests that SBS40 is the major confounder in the signature assignments obtained by the PCAWG consortium. t-tests are performed to compare the fraction shown in (b) between samples with non-zero and zero exposures of a signature. Only significant results are shown.



Extended Data Fig. 20 Details of PCAWG reanalysis for SBS signatures. a. Microsatellite mutation rate (as obtained from [70]) vs. relative exposure of MMRD-related SBS signatures (SBS6, 14, 15, 20, 21, 26, 44) for PCAWG samples. A clear separation is observed between two classes of samples, which we identify as MMRD and MMRP, respectively. SBS14 exposure alone is used to define the single sample with concurrent MMRD and POLE-exo mutations. **b.** Relative exposure of ID1 + ID2 vs. relative exposure of SBS signatures associated with POLE-exo mutations (SBS10a-d and SBS28). A clear separation is observed along the x-axis, which is used to identify tumors with POLE-exo mutations. **c.** Dependence of the selected number of signatures r and the optimal mvNMF regularization parameter $\tilde{\lambda}$ on the number of samples and the average number of SBSs.