Accurate and sensitive mutational signature analysis with MuSiCal

Hu Jin[1†], Doga C. Gulhan[1†], Benedikt Geiger[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. Here we present MuSiCal (Mutational Signature Calculator), a rigorous analytical framework with novel algorithms that solves major 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 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.

Introduction

Mutational signature analysis has emerged as a powerful tool for unveiling the mutational processes underlying somatic DNA alterations [1–3]. Analysis of a large collection of tumors profiled with whole-genome and whole-exome sequencing has led to the discovery of a number of mutational signatures consisting of single-base substitutions (SBSs), indels (IDs), and structural variations (SVs) [4–6]. Some of these signatures have been associated with specific exogenous or endogenous mutational processes [7–9], although the majority still have unknown etiologies. Early evidence shows that application of mutational signature analysis may be informative for prognosis and treatment decisions for cancer patients [10–16].

To maximize the impact of mutational signature analysis, rigorous mathematical understanding and robust algorithms are needed to increase its accuracy and interpretability. The field has seen a rapid development of computational techniques in the past decade. Popular tools such as SigProfilerExtractor [17], SignatureAnalyzer [5, 18, 19], signature.tools.lib [20, 21], and others have achieved considerable success [22]. However, several methodological difficulties still remain [3, 23]. 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. Misleading signature assignments are pervasive in the literature, as we demonstrate using the results from two recent pan-cancer studies [20, 24] (Supplementary Fig. 1). Signatures with similar spectra – such as the relatively 'flat' ones – are particularly problematic since they are more difficult to separate from each other mathematically (Supplementary Fig. 2a, b). Inaccurate signature assignment can have meaningful implications in therapeutic decision-making. For example, observation of the homologous recombination deficiency (HRD) signature SBS3 – which suggests poly(ADP-ribose) polymerase (PARP) inhibitors as a treatment option [25, 26] – in a multiple myeloma cohort has been found to be a false positive due to flawed signature assignment [11, 23, 27].

Another difficulty is in the signature discovery step (often termed *de novo* signature discovery), in which mutational signatures are directly extracted from a cohort. Numerous tools have been developed, most of which rely on

non-negative matrix factorization (NMF) [28]. But they lead to highly variable results, even from the same dataset [5, 22]. 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) [29] (Supplementary Fig. 2c). It may also confound the tissue specificity observed for some signatures [20].

Here, we present Mutational Signature Calculator (MuSiCal), a comprehensive framework that enables accurate signature assignment as well as robust and sensitive signature discovery (Fig. 1). To solve the problems described above, MuSiCal leverages several novel methods, including minimum-volume NMF (mvNMF) [30–33], likelihood-based sparse non-negative least squares (NNLS), and a data-driven approach for systematic parameter optimization and *in silico* validation. In addition, to facilitate future studies of mutational signatures in cancer, we present an improved catalog for both signatures and their assignments by reanalyzing over 2,700 genomes from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project [34]. Our new catalog of indel (ID) signatures is 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.

Results

De novo mutational signature discovery with mvNMF

Although NMF is commonly used, it has an intrinsic problem of producing non-unique solutions [35–37]. This can be illustrated by a geometric interpretation of NMF (Fig. 2a): any convex cone that encloses the data points is a valid NMF solution, and there exist infinitely many such convex 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 [30–33]. Compared to NMF, mvNMF adds a regularization term that penalizes the volume of the convex 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 (Methods). Theoretically, mvNMF is guaranteed to recover the true underlying signatures under mild assumptions satisfied by tumor somatic mutations, whereas NMF requires much stronger assumptions that are often unrealistic [33, 37, 38] (Supplementary Notes, Supplementary Fig. 3, 4). In the examples above, mvNMF accurately recovers the true signatures without distortions (Fig. 2a, b).

For a more systematic comparison, we ran NMF and mvNMF on synthetic datasets generated using tumor type-specific SBS signatures and Dirichlet-distributed exposures (Fig. 2c, 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. 1). The improvement is especially notable for relatively flat signatures. 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. 1d).

Benchmarking MuSiCal for *de novo* signature discovery

We benchmarked the performance of MuSiCal for *de novo* signature discovery against the state-of-the-art algorithm SigProfilerExtractor [17]. To recapitulate the complexities of real-life data, synthetic datasets are simulated from tumor type-specific SBS signatures and real exposure matrices produced by PCAWG [5] (Methods). To infer the number of signatures from the dataset itself, MuSiCal employs a new parameter-free method (Methods, Extended Data Fig. 2).

To evaluate the quality of *de novo* signature discovery, we use the accuracy of 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, see Supplementary Data for examples based on other tumor types). SigProfilerExtractor discovers only 3 *de novo* signatures from this dataset, whereas MuSiCal discovers 9 (Fig. 3a). Most *de novo* signatures are mixtures of a subset of the 13 true signatures, potentially due to their strong co-occurring patterns, 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. Precision, recall, and their harmonic mean (the F1 score) can subsequently be calculated. Specifically, the likelihood-based sparse NNLS is used for this matching step at different likelihood thresholds (see next section for more details), allowing the entire precision-recall curve (PRC) and the corresponding area under PRC (auPRC) to be obtained. For the Skin-Melanoma example, MuSiCal achieves an auPRC of 0.92, outperforming SigProfilerExtractor (auPRC = 0.84).

When we apply the same metric to all 25 tumor types (each with 10 independent simulations), MuSiCal achieves higher mean auPRC in 18 tumor types (9 are statistically significant) compared to SigProfilerExtractor (Fig. 3b). In the remaining 7 tumor types, MuSiCal has similar mean auPRC (2 equal, 5 lower, and none statistically significant). In Fig. 3c, we show the PRC averaged across all tumor types, and MuSiCal outperforms SigProfilerExtractor overall (auPRC of 0.929 vs. 0.893). Specifically, MuSiCal achieves higher precision at the same recall (Fig. 3d) and higher recall at the same precision (Fig. 3e). MuSiCal outperforms SigProfilerExtractor even when SigProfilerExtractor's built-in input normalization is turned on (Supplementary Fig. 5). The performance of SigProfilerExtractor is improved – although still worse than MuSiCal – when it is forced to select the same number of signatures as MuSiCal (Supplementary Fig. 5). This observation suggests that the improved performance of MuSiCal relies on both 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. 3f), and a better choice of the number of signatures, which in turn benefits from the uniqueness and stability of mvNMF solutions. Further, MuSiCal's improved performance is robust to random noise at different levels (Methods, Extended Data Fig. 3) and unknown spurious signatures present in the data (Methods, Supplementary Fig. 6). MuSiCal also significantly outperforms three additional tools (signature.tools.lib [20, 21], SignatureAnalyzer [5, 18, 19, 39], and SigneR [40]) representing different underlying approaches (Methods, Extended Data Fig. 4). Together, these results demonstrate that MuSiCal outperforms the state-of-the-art algorithms for *de novo* signature discovery in simulations emulating real-life datasets in diverse scenarios.

Refitting with likelihood-based sparse 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 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) as illustrated with a simulated example in Fig. 4a. A common method to introduce sparsity in refitting is thresholded NNLS, where signatures that contribute less than a specified percentage to the total mutation burden are set to have zero exposures [41]. The problem with a fixed threshold is that different mutational signatures require different numbers of mutations to provide confidence in their presence. Sparse signatures with prominent weights only in a few mutation channels may be statistically significant even with a small number of mutations, whereas flatter signatures may not be significant even with many mutations. Indeed, in the simulated example above, there is no fixed threshold that recovers the true signatures without adding false positives (Fig. 4a).

We propose a sparse NNLS algorithm based on multinomial likelihoods (Methods, Extended Data Fig. 5a). Briefly, we first perform NNLS to obtain an initial exposure estimate and then refine the exposure vector in a bidirectional stepwise manner. During each step, a signature is considered for addition or subtraction based on its impact on the overall multinomial likelihood of the observed mutation count vector. The use of multinomial likelihoods provides a rigorous way to assess the presence of a signature, regardless of its flatness. In the simulated example above, likelihood-based sparse NNLS is able to identify the set of active signatures correctly without any false positives if the likelihood threshold is within a proper range, alleviating the over-assignment problem (Fig. 4b, see next section on how to choose this threshold). Of note, the widely used cosine similarity is less powerful in separating similar signatures compared to multinomial likelihood (Extended Data Fig. 5b) and fails to discover the set of active signatures correctly at any threshold in the same example (Extended Data Fig. 5c). When systematically benchmarked on synthetic datasets with tumor type-specific SBS signatures (Methods), likelihood-based sparse NNLS outperforms thresholded NNLS (auPRC of 0.872 vs. 0.824) as well as two state-of-the-art algorithms – sigLASSO [42] and the Decomposition module of SigProfilerExtractor [17], achieving higher precision at the same recall and higher recall at the same precision (Fig. 4c, Supplementary Fig. 7). 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. 5d).

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, methods are lacking for systematically selecting parameters involved in post-processing and quantifying the consistency between final signature assignment and data. To tackle this, we propose a novel data-driven approach for *in silico* validation (Methods, Fig. 5a). Denote the *de novo* discovery results as $(W_{\text{data}}, H_{\text{data}})$ and the final signature assignments as $(W_{\text{s}}, H_{\text{s}})$ (s stands for sparse), which are obtained through post-processing steps from $(W_{\text{data}}, H_{\text{data}})$. We first simulate a synthetic dataset X_{simul} from $(W_{\text{s}}, H_{\text{s}})$. 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*. Compared to the difference between X_{simul} and X_{data}, which merely reflects reconstruction errors, the difference between $(W_{\text{simul}}, H_{\text{simul}})$ and $(W_{\text{data}}, H_{\text{data}})$ captures more intricate structures of the dataset as learned through the *de novo* discovery step. 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 [5] (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. 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 [5].

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, Supplementary Fig. 8).

PCAWG reanalysis with MuSiCal

We reanalyzed over 2,700 cancer genomes from PCAWG for both SBS and indel (ID) signatures to derive a refined catalog 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 (Methods, Extended Data Fig. 6). Samples with mismatch repair deficiency (MMRD) and/or polymerase epsilon exonuclease (POLE-exo) domain mutations are isolated with a dedicated procedure beforehand and analyzed separately in a tissue-independent manner (Methods).

ID signatures have been under-studied until recently, in part due to the relatively low mutational burden of IDs compared to SBSs [5, 7, 9, 43–45]. Our reanalysis results in a substantial revision and expansion of the COSMIC catalog of ID signatures (Fig. 6a). Among the 18 ID signatures in the current COSMIC catalog, 16 are discovered with modified spectra in our reanalysis (ID1-14, 17, and 18). The other two (ID15 and 16) are not discovered and likely reflect artifactual indels upon further inspection (Supplementary Notes, Extended Data Fig. 7a-d). In addition, our reanalysis discovers 9 new ID signatures. 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, and importantly, our updated MuSiCal version of ID4 is more similar to the experimentally derived signature of topoisomerase 1 (TOP1)-mediated deletions [43] (Supplementary Notes, Extended Data Fig. 7e-f).

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 (potentially benefiting from a separate analysis for them), likely representing nonlinear interactions between DNA damage and repair [46] (Extended Data Fig. 8). 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%) (Supplementary Fig. 9). 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 [5] (Fig. 6b, Supplementary Fig. 10, Supplementary Table 3 and 4).

SBS-ID associations 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 (Fig. 6c), 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 [7, 9], co-occurrence of ID and SBS signatures can indicate a common origin.

Apart from expected correlations between ID and SBS signatures with known and related etiologies (Supplementary Notes), novel correlations are observed for both newly added ID signatures in our updated catalog and existing signatures with revised spectra (Fig. 6c). ID11b – unlike COSMIC ID11 – correlates with both SBS (SBS4 and SBS92 [17]) and ID (ID3) signatures known to be associated with tobacco smoking (Fig. 6c, Extended Data Fig. 9a, 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. 9c) [7]. 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. 9c) [7]. Interestingly, all four chemicals produce SBS signatures similar to SBS4 (Extended Data Fig. 9c) [7]. In smoking-related (i.e., SBS4- or SBS92-positive) lung, liver, head-and-neck, and bladder cancers, ID11b is indeed required to explain the observed abundance of 1-bp T insertions following a T (Extended Data Fig. 9d). 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 ($\geq$ 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 (Supplementary Fig. 11).

Improved signature assignment resolves ambiguities with flat signatures

To demonstrate the improved signature assignments by MuSiCal (Fig. 6b, Supplementary Fig. 10), we focused 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. Thus distinguishing SBS40 from other signatures and quantifying its contribution in different samples are mathematically difficult. PCAWG assigns SBS40 to nearly every tumor type (Supplementary Fig. 10) [5]. However, we believe this result is artifactual, caused by the over-assignment problem described earlier. Some studies have noticed this problem and have manually combined SBS40 with other flat signatures [47, 48]. 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 Supplementary Fig. 12) 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 [5] 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 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. 10). 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 and thus confounds the interpretation of these signatures. For example, signature assignments from the PCAWG consortium show distinct ratios between the exposures of SBS1 and SBS5 (clock-like signatures) in samples with and without SBS40, even when tumor type-specific accumulation rates are factored out (Supplementary Notes, Fig. 7c, d, Supplementary Fig. 13). By comparison, this confounding effect is largely alleviated in MuSiCal-derived signature assignments (Fig. 7c, d, Supplementary Fig. 13).

Discussion

Mutational signature analysis has become routine in cancer genome analysis [49]. In this work, we developed MuSiCal that outperforms state-of-the-art algorithms for both signature discovery and assignment. MuSiCal's improved performance is further demonstrated by the improved consistency of MuSiCal-derived signature assignments with biological ground truth in real data when such ground truth is known, as is the case for HRD-associated SBS3 (Supplementary Fig. 14) and platinum-associated SBS31/35 (Supplementary Fig. 15). Although MuSiCal solves several problems in the standard workflow, other methods that perform specific tasks – such as joint discovery of different signature types [50], incorporation of epigenetic data [51], and utilization of clinical information [52] – are complementary and valuable additions to the signature analysis toolbox. There are also many other outstanding methodological challenges requiring future developments, which could potentially benefit from recent advances in machine learning and artificial intelligence [53].

The number of whole genomes of cancer and other diseases continues to grow rapidly, especially by consortium projects such as Genomics England [21, 54] and Hartwig Medical Foundation's metastatic tumor project [55, 56]. 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.

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 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 (Supplementary Fig. 9) – 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), which lack more detailed context information. As a result, some of the identified signatures may correspond to multiple mutational processes, thus leading to uncertainties in the interpretation of some signature assignment results (Supplementary Fig. 16). Application of MuSiCal to alternative signature definitions should further disentangle mutational processes and help gain insights into signatures with unknown etiologies. Finally, signatures of other mutation types, including copy number variations and structural

variations, are less well studied [6, 57–59]. We expect that MuSiCal will facilitate future developments of these signatures when their proper definitions are developed.

Acknowledgments. This work was funded by grants from the National Institutes of Health (R01CA269805 to P.J.P.), Ludwig Center at Harvard (P.J.P.), Cancer Research UK Grand Challenge and the Mark Foundation to the SPECIFICANCER team (P.J.P.), and Swedish Research Council (2020-00583 to V.L.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We would like to thank D. Glodzik for careful reading of the manuscript. We gratefully acknowledge the contributions of the many clinical networks across ICGC, TCGA and other consortia, who provided samples and data. We thank the patients and their families for their participation in the individual projects.

Author contributions. H.J., D.C.G, and P.J.P. conceived the study. P.J.P. supervised the study. H.J. and D.C.G developed the algorithms and carried out data analysis, with assistance from B.G., D.B.-I., D.G., and V.L. In particular, B.G. helped with implementing the algorithms; B.G. and V.L. helped with software testing; D.B.-I. helped with early tests of the mvNMF algorithm; D.G. helped with early tests of the *in silico* validation algorithm. H.J., D.C.G., and P.J.P. wrote the manuscript with input from the other authors. All authors read and approved the final manuscript.

Competing interests. The authors declare no competing interests.

Figure legends

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. Solid purple arrows represent a typical workflow of mutational signature analysis. Dashed orange arrows represent that some steps in the workflow potentially require repeating an earlier step (see Results).

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 (represented by the three axes x, y, and z). 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 $x + y + z = 1$. **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. The number in the parenthesis after each signature name indicates the number of tumor types where the corresponding signature is present. The number of data points in each box plot equals this number multiplied by 20 (independent simulations). 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 (marked by gray signature labels), the cosine errors for both algorithms are much smaller than 0.001 in those cases and thus negligible. Note that the apparent similar or larger spread of cosine errors for mvNMF is due to the log scale on y-axis. Solutions from mvNMF are in fact more stable, producing smaller standard deviations in the cosine errors overall. In all panels, random exposures are 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 [5] (Supplementary Fig. 3).

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 (percentages below signature names denote exposure strengths) and the exposure matrix produced by the PCAWG consortium. MuSiCal and SigProfilerExtractor are applied to derive *de novo* signatures, which are subsequently decomposed as non-negative mixtures of COSMIC signatures with likelihood-based sparse NNLS at different likelihood thresholds (see next section for more details). Precision and recall are then calculated at each threshold by comparing matched COSMIC signatures with the 13 true signatures, and auPRC is obtained. The matching result corresponding to the largest achieved F1 score is shown in the heatmap on the right. **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 auPRC for each individual dataset. Tumor types are sorted according to the mean auPRC gain by MuSiCal. *: $p < 0.05$. **: $p < 0.005$. ***: $p < 0.0005$. p-values are calculated with two-sided paired t-tests. Raw p-values from left to right (significant ones only): 7.5×10^{-10}, 0.0034, 8.4×10^{-5}, 1.0×10^{-4}, 8.3×10^{-5}, 5.6×10^{-5}, 0.0012, 0.044, 0.032, 0.012. **c.** PRC of MuSiCal and SigProfilerExtractor averaged across all tumor types. **d.** Precision of MuSiCal and SigProfilerExtractor averaged across all tumor types. Recall is fixed at 0.9. Error bars indicate standard deviation from 10 independent simulations. **e.** Recall of MuSiCal and SigProfilerExtractor averaged across all tumor types. Precision is fixed at 0.98, corresponding to an FDR of 2%. Error bars indicate standard deviation from 10 independent simulations. **f.** Cosine reconstruction errors for MuSiCal- and SigProfilerExtractor-derived *de novo* signatures ($n = 1798$ and 1564, respectively). 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-value is calculated with a two-sided t-test.

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.** Performance of likelihood-based sparse NNLS and three other algorithms for refitting on simulated datasets (Methods). For MuSiCal and thresholded NNLS, the entire precision-recall curve (PRC) is shown by scanning the corresponding threshold. For SigProfilerExtractor Decomposition and sigLASSO, a PRC can not be generated since multiple parameters are involved in each algorithm, and thus the precision and recall achieved with default parameters are shown.

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}) [5] 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.

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 mismatch repair proficient (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.

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 two-sided 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) (Supplementary Fig. 13a). Error bands (shaded areas) indicate 95% confidence intervals of the linear regressions. **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.

References

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

[2] 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).

[3] 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).

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

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

[6] 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.

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

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

[9] 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).

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

[11] 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).

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

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

[14] 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).

[15] 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).

[16] 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).

[17] Islam, S. A. *et al.* Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics* 100179 (2022).

[18] Kasar, S. *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nature Communications* **6**, 8866 (2015).

[19] Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biol* **20** (1), 228 (2019).

[20] 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).

[21] Degasperi, A. *et al.* Substitution mutational signatures in whole-genome-sequenced cancers in the UK population. *Science* **376** (6591) (2022).

[22] 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).

[23] 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.

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

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

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

[27] 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).

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

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

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

[31] 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).

[32] 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).

[33] 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).

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

[35] 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).

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

[37] 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).

[38] 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).

[39] Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genet* **48** (6), 600–606 (2016).

[40] Rosales, R. A., Drummond, R. D., Valieris, R., Dias-Neto, E. & da Silva, I. T. signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics (Oxford, England)* **33** (1), 8–16 (2017).

[41] 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).

[42] 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).

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

[44] 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).

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

[46] 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).

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

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

[49] 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).

[50] Funnell, T. *et al.* Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. *PLoS Comput Biol* **15** (2), e1006799 (2019).

[51] Vöhringer, H., Hoeck, A. V., Cuppen, E. & Gerstung, M. Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. *Nat Commun* **12** (1), 3628 (2021).

[52] Afsari, B. *et al.* Supervised mutational signatures for obesity and other tissue-specific etiological factors in cancer. *Elife* **10** (2021).

[53] Sanjaya, P. *et al.* Mutation-Attention (MuAt): deep representation learning of somatic mutations for tumour typing and subtyping. *Genome Med* **15** (1), 47 (2023).

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

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

[56] Pich, O. *et al.* The mutational footprints of cancer therapies. *Nat Genet* **51** (12), 1732–1740 (2019).

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

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

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

Methods

Ethical Approval

This study is based on reanalysis of previously published data and complies with ethical regulations within the respective cohorts (TCGA and ICGC).

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.

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 [30–33]. In mvNMF, the following optimization problem is solved [33],

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

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 [33] 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 [28]. 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 [33]. 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 number of samples or mutations and thus more generalizable across different datasets, 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 [32]. 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 [60]. A solution is said to be significantly worse if either of the two tests gives a p-value smaller than 0.05. This procedure ensures that the regularization does not significantly sacrifice reconstruction accuracy, while still promoting uniqueness of the solutions (Supplementary Fig. 17).

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.

2). 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 [2, 17, 20]. 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. 2c). 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 [61, 62]. 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. 2g, h). We thus select the largest r among those with $k = r$ as the inferred number of signatures (Extended Data Fig. 2b). 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.

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. 5a). 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.

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 (Supplementary Fig. 8). 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 [60] 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 (Supplementary Fig. 8): 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.

Preprocessing

MuSiCal provides two complementary methods for preprocessing in order to improve the sensitivity of *de novo* signature discovery (Extended Data Fig. 6).

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. 6a-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 [61, 62]. 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. 6d-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, 0.05).

Computational cost

For *de novo* signature discovery, MuSiCal (with mvNMF) requires similar but slightly more computational time compared to SigProfilerExtractor (with NMF), and considerably less memory (Supplementary Fig. 18). The computational costs of matching and refitting steps in MuSiCal are negligible compared to *de novo* signature discovery. The *in silico* parameter optimization step requires rerunning *de novo* signature discovery for a grid of thresholds. But during this grid search, there is no need to select the regularization parameter in mvNMF or the number of signatures (as they are both fixed), which are the most time-consuming calculations. Thus, the computational cost for *in silico* optimization is also small compared to *de novo* signature discovery when parallelized properly.

Simulation studies

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. 1), 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 [5]. 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 [5] (Supplementary Fig. 3). 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 [28]. Simulation was performed with the function `simulate_LDA()` from the simulation module in MuSiCal (`musical.simulation`).

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, Supplementary Fig. 5), 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 [5] to simulate mutation count matrices through multinomial sampling (using the function `simulate_count_matrix()` from `musical.utils`). 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 [5] 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`) except in the results denoted by SigProfilerExtractor-norm in Supplementary Fig. 5, where it was turned on with `-nx gmm`. For MuSiCal, the preprocessing module was skipped. In Extended Data Fig. 3, we added random Gaussian noise at different levels (1%, 2.5%, 5%, and 10%) to the simulated genomes described above (which already included random Poisson sampling noises). In more detail, we followed [17] and resampled each element of the mutation count matrix from a Gaussian distribution where the mean was equal to the original count and the standard deviation was equal to the original count multiplied by the noise level. The resulting matrix was then rounded to the closest integers, and negative values were set to 0. In Supplementary Fig. 6, we spiked in mutations from COSMIC SBS48 to the simulated genomes. The number of spike-in mutations was equal to 5% of the mutational burden for each sample. SBS48 was chosen to represent a spurious signature because it is not present in the original synthetic dataset and is annotated in COSMIC as a possible sequencing artifact found in cancer samples that were subsequently blacklisted for poor quality of sequencing data. To pretend that SBS48 was unknown, we removed SBS48 from the COSMIC catalog when matching the *de novo* signatures and calculating precision and recall. The auPRC scores were calculated with `sklearn.metrics.average_precision_score()` from scikit-learn version 0.24.1 [63], which is more accurate than using the trapezoidal rule.

Simulation studies benchmarking other existing tools for *de novo* signature discovery

We benchmarked the performance of MuSiCal for *de novo* signature discovery mainly against SigProfilerExtractor because it has been the most popular tool in the field and was shown to outperform a number of other existing tools in a recent benchmark study [17]. COSMIC signatures and the PCAWG study also largely relied on SigProfilerExtractor. To be more comprehensive in the benchmark analysis (Extended Data Fig. 4), we selected three additional tools based on the consideration that different underlying algorithms should be represented and popular tools (especially those used in large landmark studies) should be included. Specifically, signature.tools.lib [20, 21] is based on an R implementation of NMF (SigProfilerExtractor is based on a Python implementation of NMF), included several methodological improvements [20], and was used in the recent landmark study of the Genomics England Cohort [21]; SignatureAnalyzer [18, 19, 39] is based on a Python implementation of Bayesian NMF and was used in the PCAWG

study as well [5]; and SigneR is based on an R implementation of Bayesian NMF [40]. Compared to the four included in our benchmark analysis, most other existing tools only differ in implementation details or specific functionalities [17, 22].

For signature.tools.lib, version 2.1.2 was used with recommended parameters (`nrepeats=200, nboots=20, clusteringMethod="MC", filterBestOfEachBootstrap=TRUE, filterBest_RTOL=0.001, filterBest_nmaxtokeep=10, nmfmethod="brunet"`). As signature.tools.lib does not select the optimal number of signatures automatically, we followed the recommendation described in https://github.com/Nik-Zainal-Group/signature.tools.lib and performed manual selection based on `norm.Error (orig. cat.)` and `Ave.SilWid.MC` by visually inspecting the produced plots. For SignatureAnalyzer, version 0.0.7 was used with default parameters, except that more iterations than default were used (`-n 20`). For SigneR, version 1.22.0 was used with default parameters.

Simulation studies benchmarking different refitting algorithms

To benchmark the performance of different algorithms for refitting (Fig. 4c, Supplementary Fig. 7), 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. 1, using the function `simulate_LDA()` from the simulation module in MuSiCal (`musical.simulation`). 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 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. In Fig. 4c and Supplementary Fig. 7a-c, precision and recall were calculated by comparing the zero and nonzero entries between the true exposure matrix and the exposure matrix obtained from refitting. The auPRC scores were calculated with `sklearn.metrics.average_precision_score()` from scikit-learn version 0.24.1 [63], which is more accurate than using the trapezoidal rule. For likelihood-based sparse NNLS and thresholded NNLS (also implemented in MuSiCal) in Supplementary Fig. 7d, e, the optimal threshold was chosen through a grid search to achieve the maximum correct support discovery rate when averaged across the 10 independent simulations.

Reanalysis of PCAWG data

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 [64] (Supplementary Fig. 19a, 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) (Supplementary Fig. 19a, 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 Supplementary Fig. 19c.

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 2.3% of samples (ranging between 0% to 7.0% in different tumor types). For Kidney.RCC, Liver.HCC, and Skin.Melanoma, we further stratified samples into subsets with distinct signature compositions using hierarchical clustering on the exposure matrix H obtained from the initial *de novo* signature discovery. The distinct clusters within these tumor types were defined by several hypermutator processes including aristolochic acid, aflatoxin, and UV exposures. 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. 6a-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.

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.

Statistics and Reproducibility

This study was designed to be a retrospective analysis of previously published data. No statistical method was used to predetermine sample size. The Investigators were not blinded to allocation during experiments and outcome assessment. For *de novo* signature discovery in the PCAWG reanalysis, samples with low SBS or indel counts were

excluded, as described in detail above. All box plots indicate median (center line), upper and lower quartiles (box limits), and 1.5x interquartile range (whiskers).

Data availability

PCAWG data, including mutation count matrices of SBSs and IDs, as well as exposure matrices obtained by [5], 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/. Source data, including MuSiCal-derived signature catalog from the PCAWG reanalysis, are provided in Supplementary Tables 1-4.

Code availability

MuSiCal is implemented in Python and available at https://github.com/parklab/MuSiCal [65]. SigProfilerExtractor (version 1.1.3) was downloaded from https://github.com/AlexandrovLab/SigProfilerExtractor. sigLASSO (version 1.1) was downloaded from https://github.com/gersteinlab/siglasso. signature.tools.lib (version 2.1.2) was downloaded from https://github.com/Nik-Zainal-Group/signature.tools.lib. SignatureAnalyzer (version 0.0.7) was downloaded from https://github.com/getzlab/SignatureAnalyzer. SigneR (version 1.22.0) was obtained from https://www.bioconductor.org/packages/release/bioc/html/signeR.html. Custom scripts and analysis notebooks for reproducing results in the paper are available on Zenodo [66].

References

[60] 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).

[61] 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).

[62] 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).

[63] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

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

[65] Jin, H., Gulhan, D. C., Geiger, B. & Ljungstrom, V. parklab/MuSiCal: v1.0.0 (2023). URL https://doi.org/10.5281/zenodo.10257728.

[66] Jin, H. & Gulhan, D. C. Analysis code for MuSiCal paper (2023). URL https://doi.org/10.5281/zenodo.10291569.

Figures

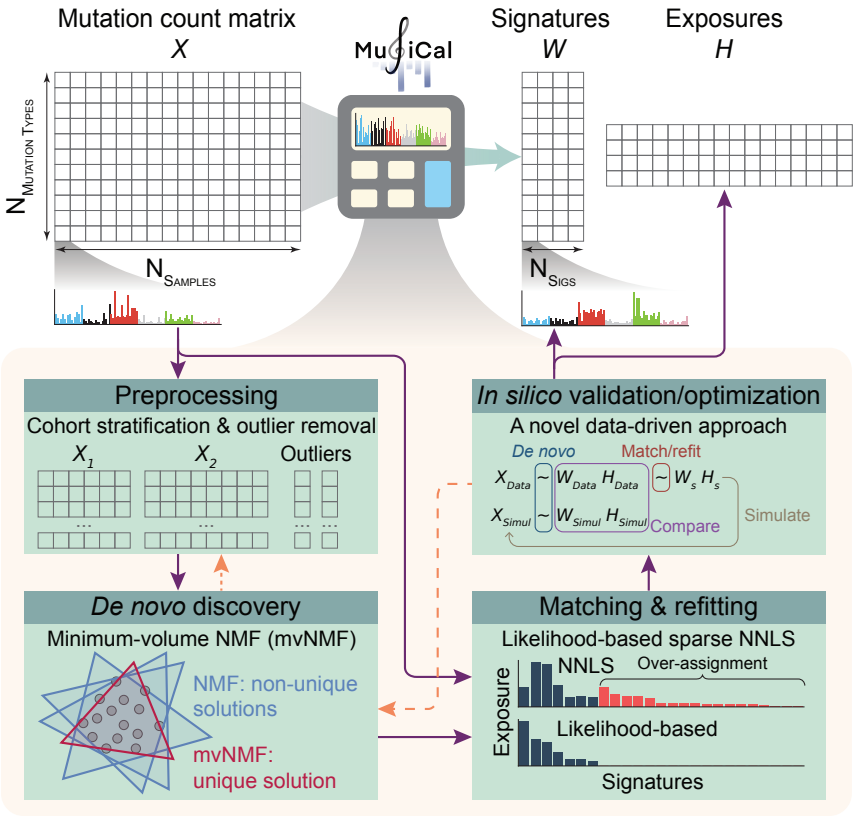


Fig. 1

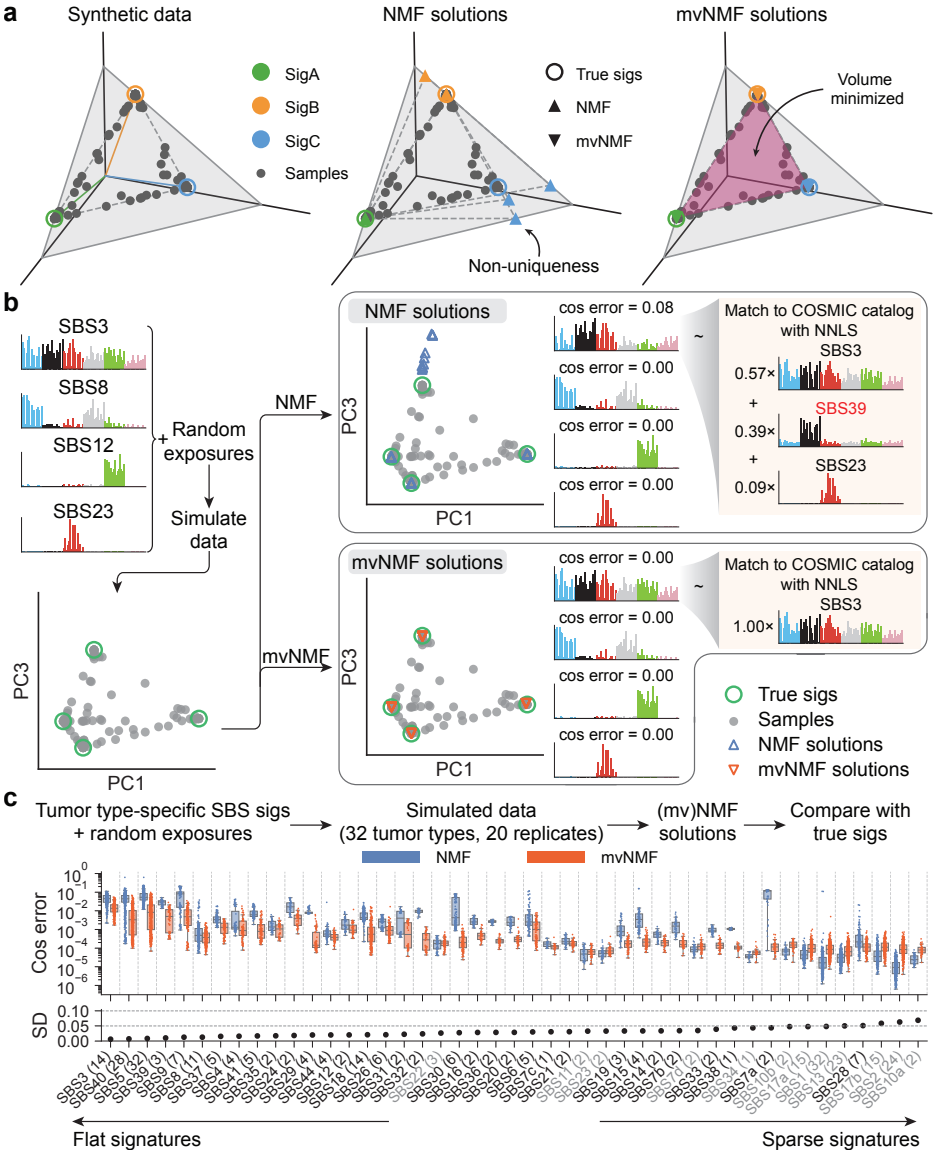

a
Synthetic data
NMF solutions
mvNMF solutions
SigA
SigB
SigC
Samples
True sigs
NMF
mvNMF
Volume minimized
Non-uniqueness
b
SBS3
SBS8
SBS12
SBS23
Random exposures
Simulate data
NMF
mvNMF
PC3
PC1
NMF solutions
mvNMF solutions
cos error = 0.08
cos error = 0.00
Match to COSMIC catalog with NNLS
0.57×
0.39×
0.09×
1.00×
SBS39
True sigs
Samples
NMF solutions
mvNMF solutions
c
Tumor type-specific SBS sigs + random exposures
Simulated data (32 tumor types, 20 replicates)
(mv)NMF solutions
Compare with true sigs
NMF
mvNMF
Cos error
SD
Flat signatures
Sparse signatures

Fig. 2

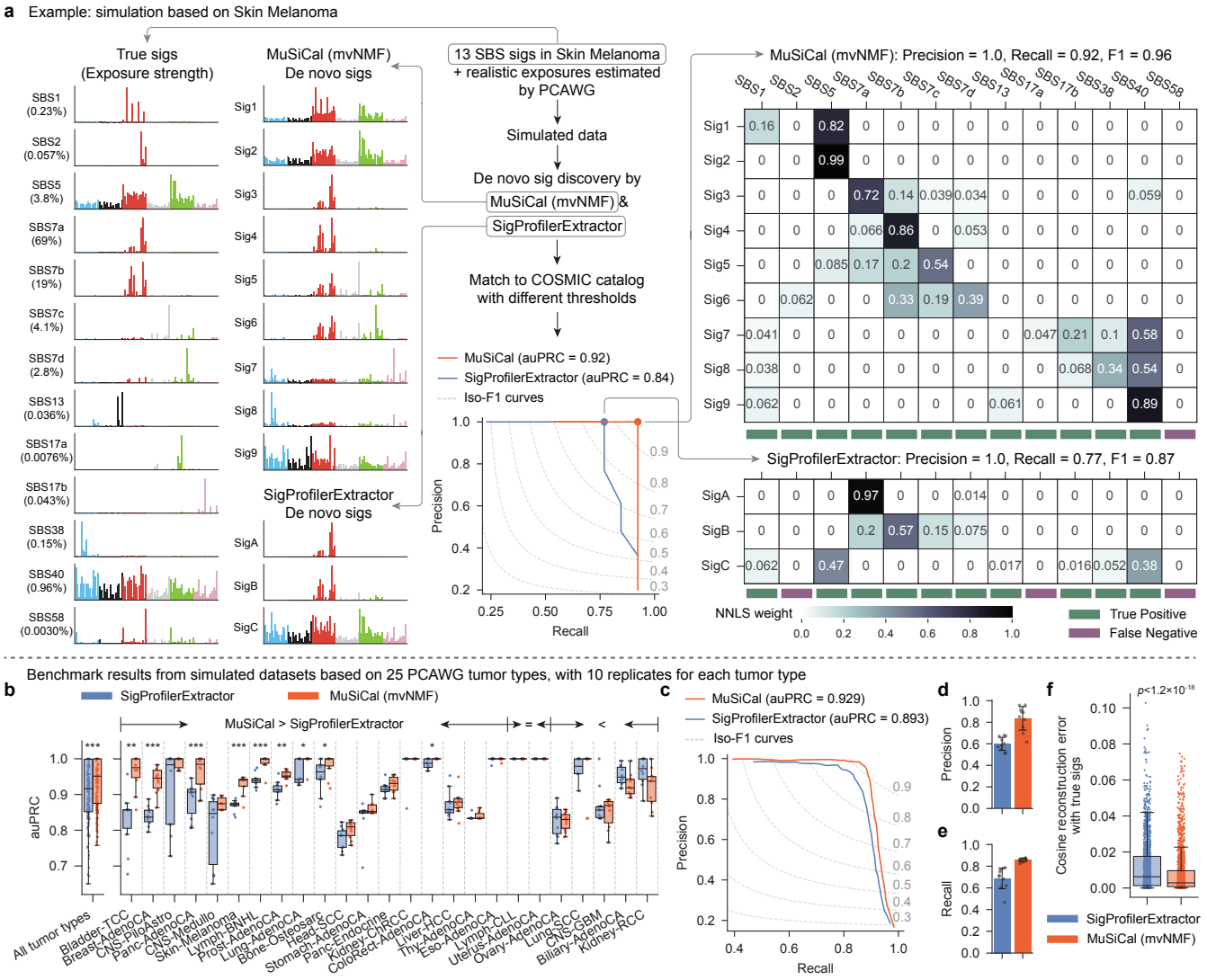

a Example: simulation based on Skin Melanoma
True sigs (Exposure strength)
MuSiCal (mvNMF) De novo sigs
SigProfilerExtractor De novo sigs
13 SBS sigs in Skin Melanoma + realistic exposures estimated by PCAWG
Simulated data
De novo sig discovery by
MuSiCal (mvNMF) &
SigProfilerExtractor
Match to COSMIC catalog with different thresholds
MuSiCal (auPRC = 0.92)
SigProfilerExtractor (auPRC = 0.84)
Iso-F1 curves
Precision
Recall
MuSiCal (mvNMF): Precision = 1.0, Recall = 0.92, F1 = 0.96
SigProfilerExtractor: Precision = 1.0, Recall = 0.77, F1 = 0.87
NNLS weight
True Positive
False Negative
b Benchmark results from simulated datasets based on 25 PCAWG tumor types, with 10 replicates for each tumor type
SigProfilerExtractor
MuSiCal (mvNMF)
MuSiCal > SigProfilerExtractor
auPRC
All tumor types
c
MuSiCal (auPRC = 0.929)
SigProfilerExtractor (auPRC = 0.893)
Iso-F1 curves
d
e
f
Cosine reconstruction error with true sigs
SigProfilerExtractor
MuSiCal (mvNMF)

Fig. 3

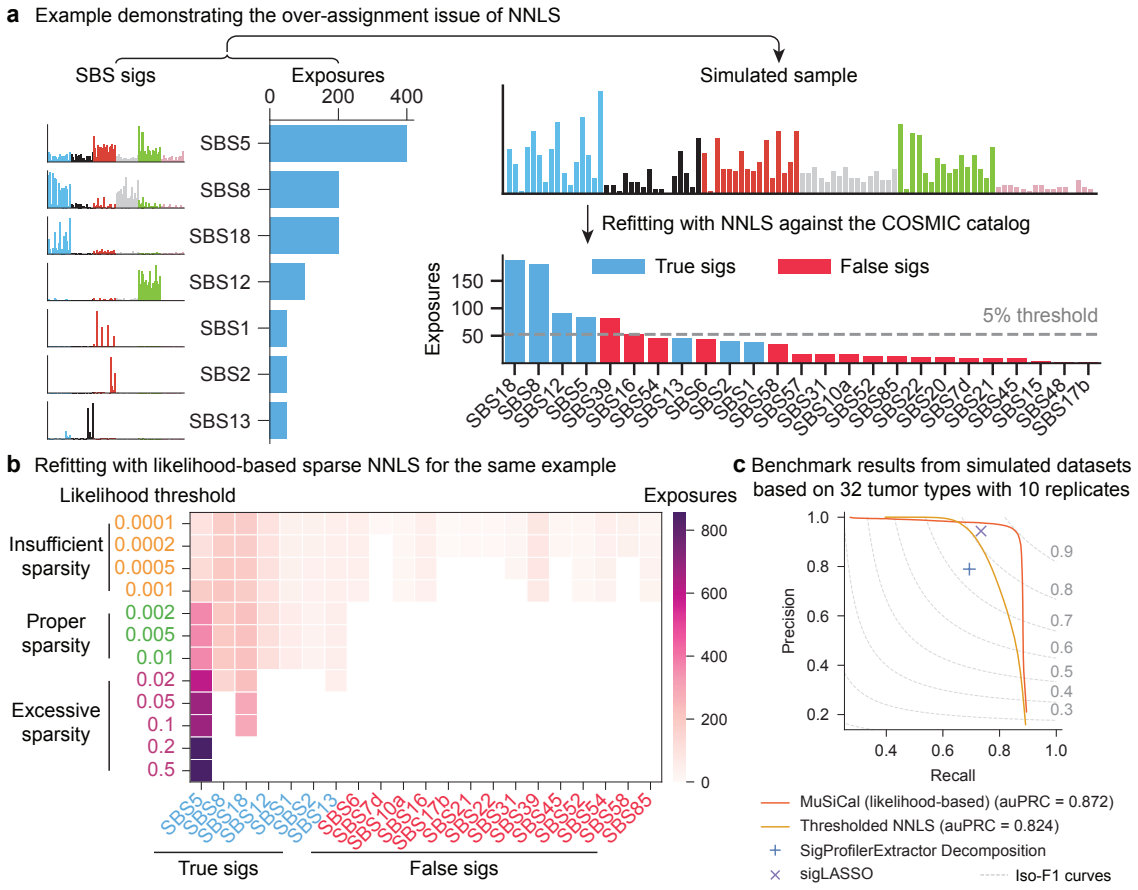

a Example demonstrating the over-assignment issue of NNLS
SBS sigs
Exposures
0
200
400
SBS5
SBS8
SBS18
SBS12
SBS1
SBS2
SBS13
Simulated sample
Refitting with NNLS against the COSMIC catalog
True sigs
False sigs
Exposures
150
100
50
5% threshold
SBS18
SBS8
SBS12
SBS5
SBS39
SBS16
SBS54
SBS13
SBS6
SBS2
SBS1
SBS58
SBS57
SBS31
SBS10a
SBS52
SBS85
SBS22
SBS20
SBS7d
SBS21
SBS45
SBS15
SBS48
SBS17b
b Refitting with likelihood-based sparse NNLS for the same example
Likelihood threshold
Exposures
Insufficient sparsity
Proper sparsity
Excessive sparsity
0.0001
0.0002
0.0005
0.001
0.002
0.005
0.01
0.02
0.05
0.1
0.2
0.5
800
600
400
200
0
True sigs
False sigs
c Benchmark results from simulated datasets based on 32 tumor types with 10 replicates
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.9
0.8
0.7
0.6
0.5
0.4
0.3
MuSiCal (likelihood-based) (auPRC = 0.872)
Thresholded NNLS (auPRC = 0.824)
SigProfilerExtractor Decomposition
sigLASSO
Iso-F1 curves

Fig. 4

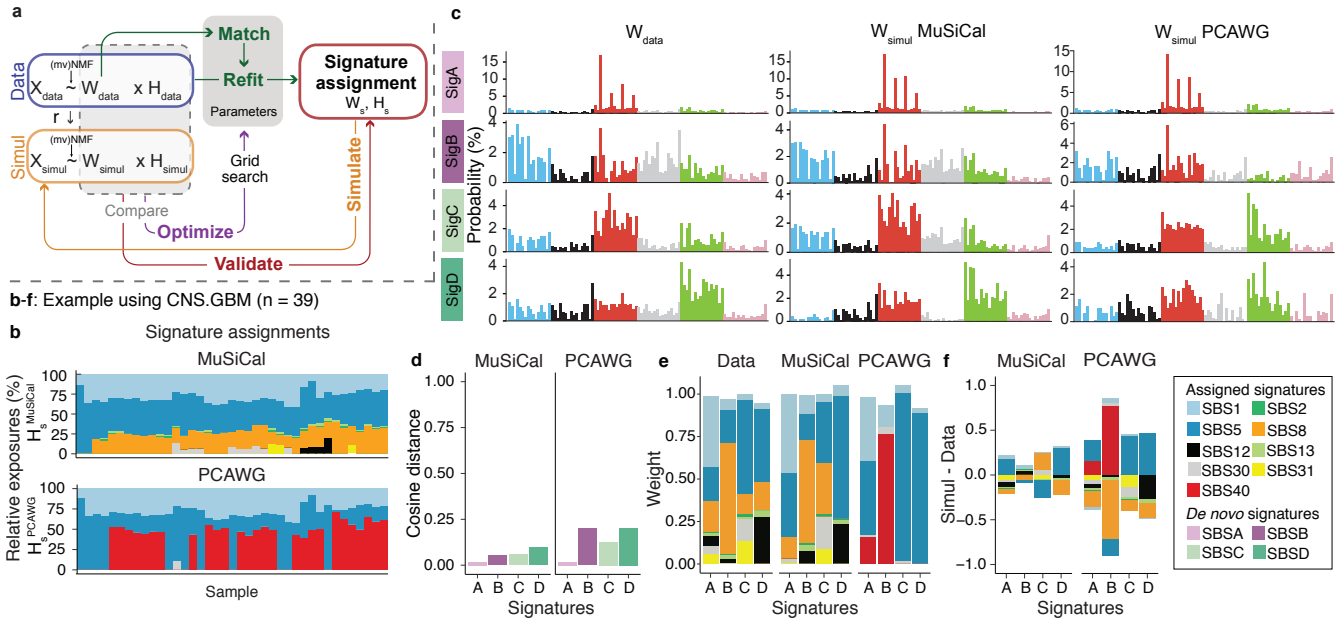

a
Data
Simul
(mv)NMF
Match
Refit
Parameters
Signature assignment
Grid search
Compare
Optimize
Simulate
Validate
b-f: Example using CNS.GBM (n = 39)
b
Signature assignments
MuSiCal
PCAWG
Relative exposures (%)
Sample
c
SigA
SigB
SigC
SigD
Probability (%)
MuSiCal
d
Cosine distance
Signatures
e
Data
Weight
f
Simul - Data
Assigned signatures
SBS1
SBS2
SBS5
SBS8
SBS12
SBS13
SBS30
SBS31
SBS40
De novo signatures
SBSA
SBSB
SBSC
SBSD

Fig. 5

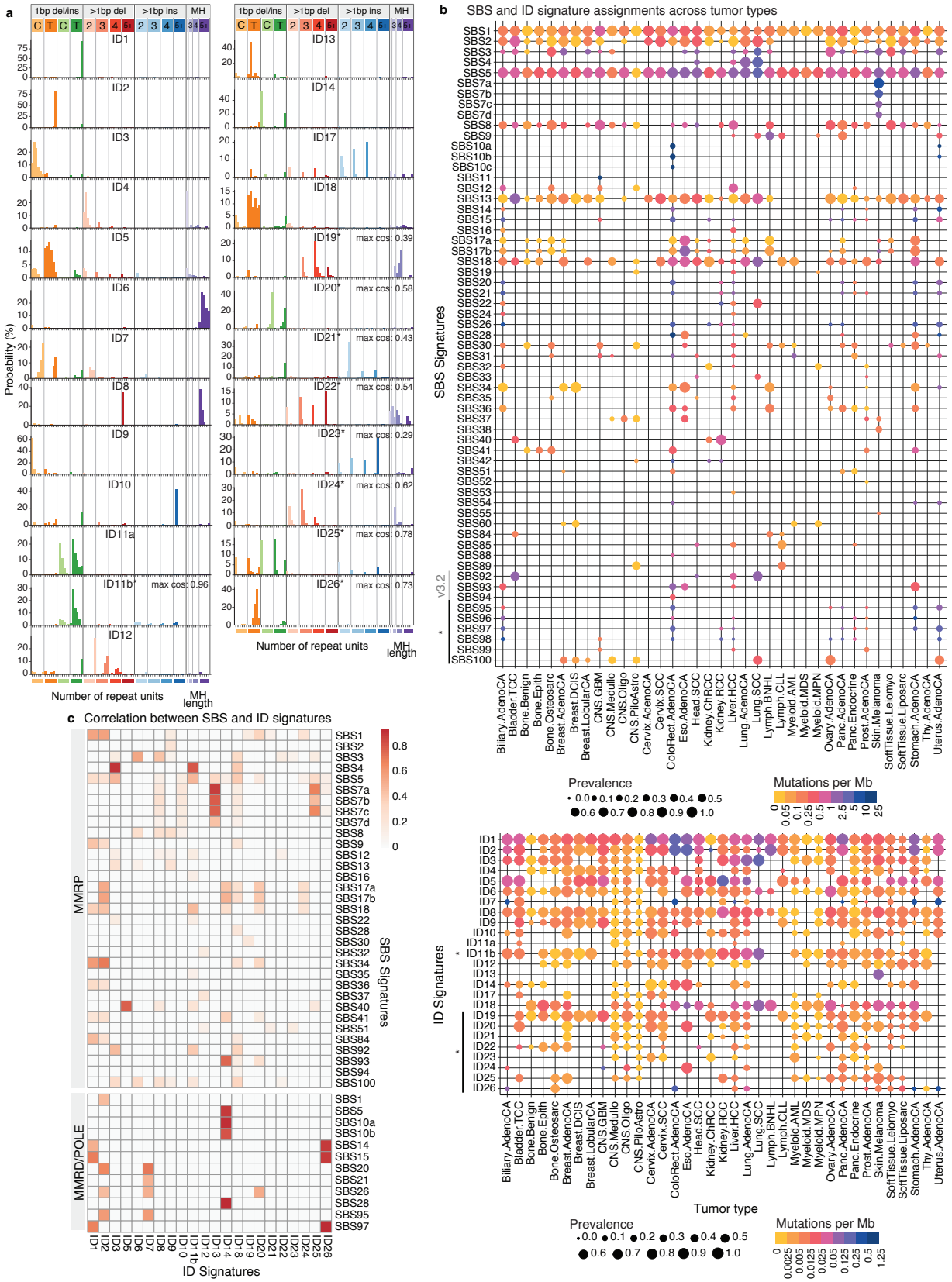

a
1bp del/ins
>1bp del
>1bp ins
MH
Probability (%)
Number of repeat units
MH length
b
SBS and ID signature assignments across tumor types
SBS Signatures
v3.2
Prevalence
Mutations per Mb
ID Signatures
Tumor type
c
Correlation between SBS and ID signatures
MMRP
MMRD/POLE
SBS Signatures
ID Signatures

Fig. 6

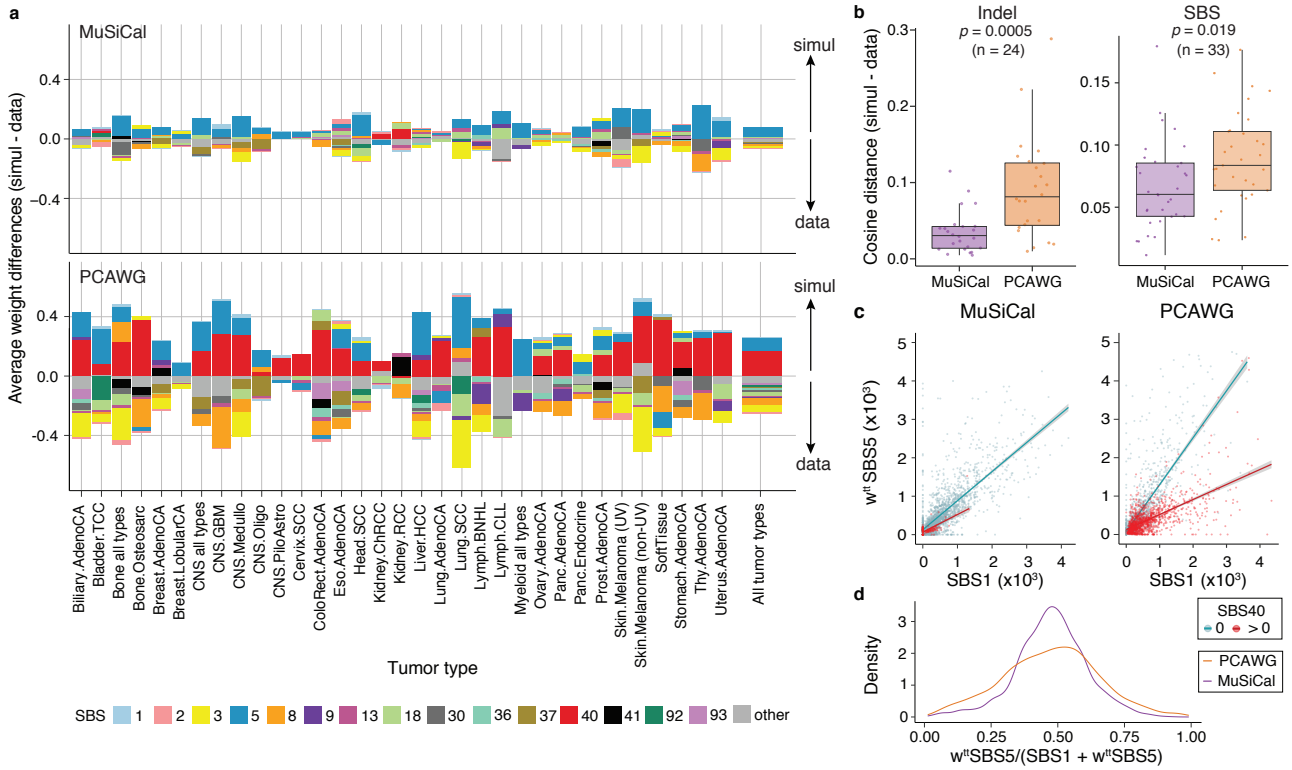
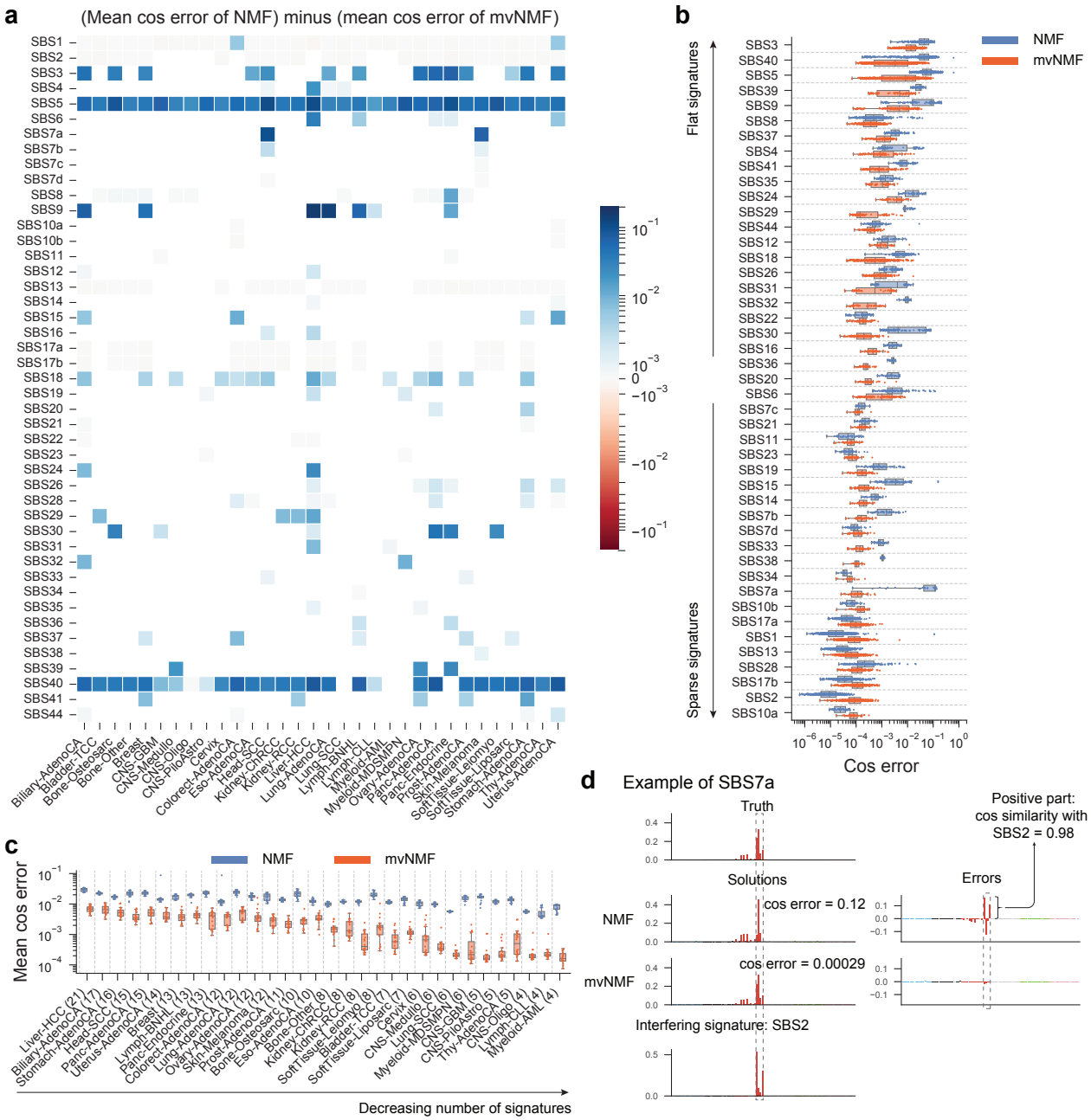
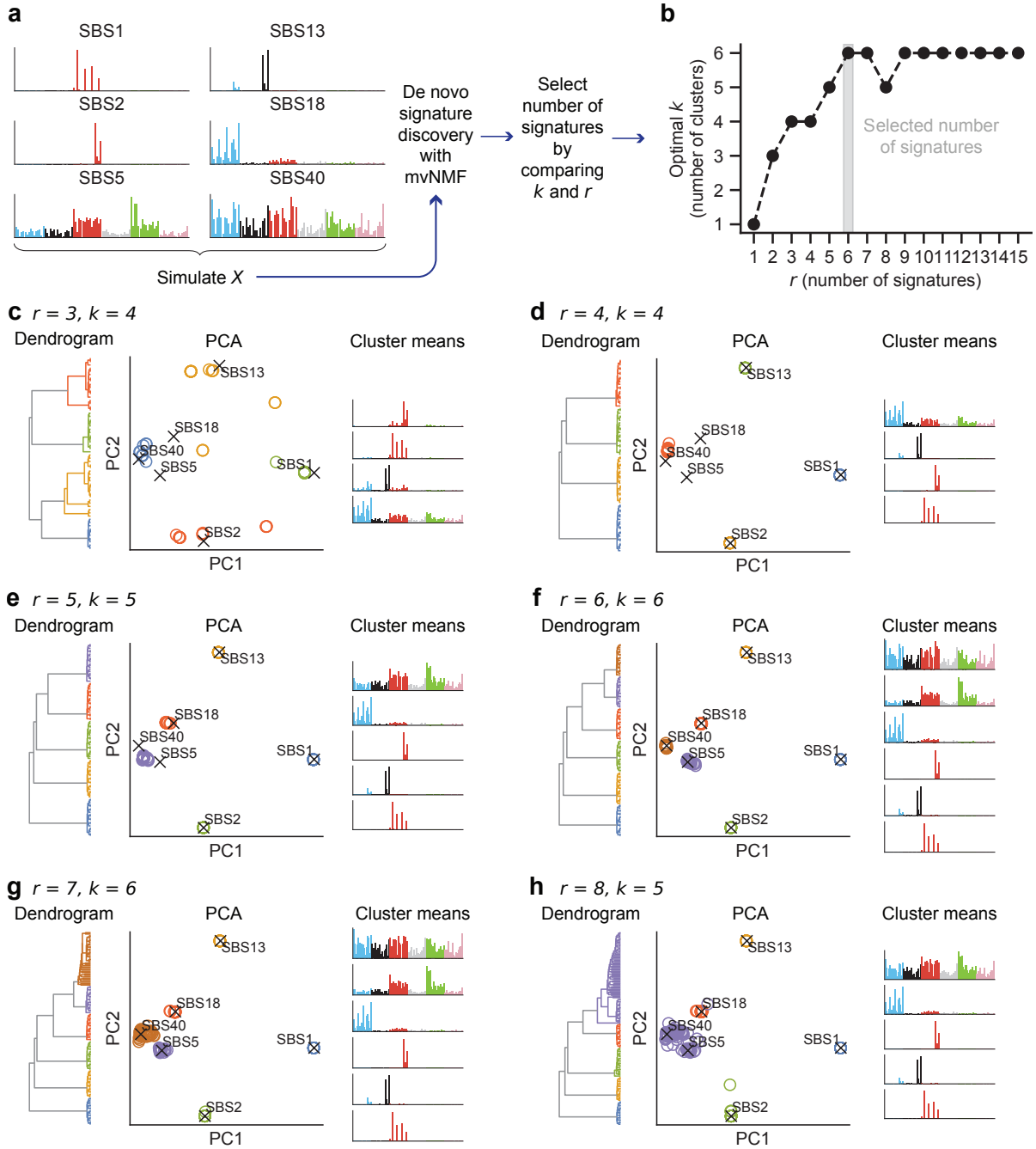

a
MuSiCal
PCAWG
simul
data
Average weight differences (simul - data)
Tumor type
Biliary.AdenoCA
Bladder.TCC
Bone all types
Bone.Osteosarc
Breast.AdenoCA
Breast.LobularCA
CNS all types
CNS.GBM
CNS.Medullo
CNS.Oligo
CNS.PiloAstro
Cervix.SCC
ColoRect.AdenoCA
Eso.AdenoCA
Head.SCC
Kidney.ChRCC
Kidney.RCC
Liver.HCC
Lung.AdenoCA
Lung.SCC
Lymph.BNHL
Lymph.CLL
Myeloid all types
Ovary.AdenoCA
Panc.AdenoCA
Panc.Endocrine
Prost.AdenoCA
Skin.Melanoma (UV)
Skin.Melanoma (non-UV)
SoftTissue
Stomach.AdenoCA
Thy.AdenoCA
Uterus.AdenoCA
All tumor types
SBS 1 2 3 5 8 9 13 18 30 36 37 40 41 92 93 other
b
Indel
p = 0.0005
(n = 24)
SBS
p = 0.019
(n = 33)
Cosine distance (simul - data)
MuSiCal
PCAWG
c
MuSiCal
PCAWG
wtt SBS5 (x10³)
SBS1 (x10³)
d
Density
wttSBS5/(SBS1 + wttSBS5)
SBS40
0
> 0
PCAWG
MuSiCal

Fig. 7

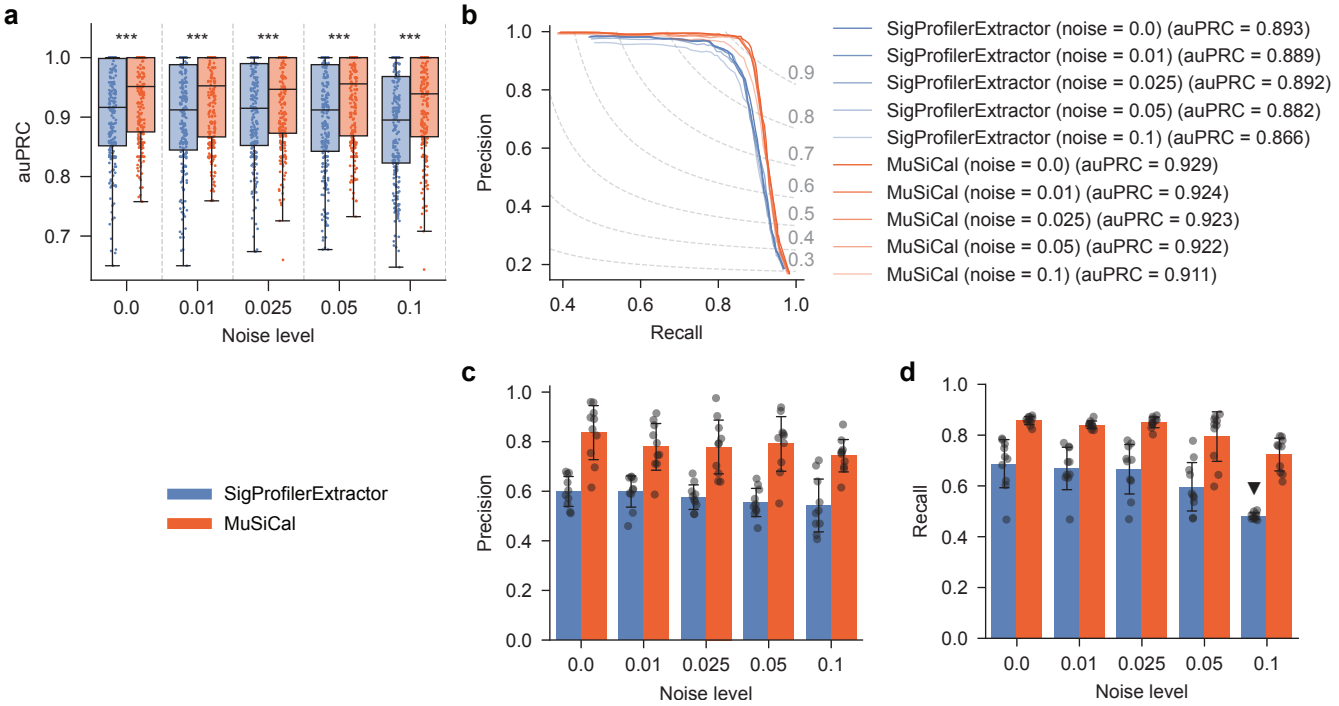
Extended Data Figures



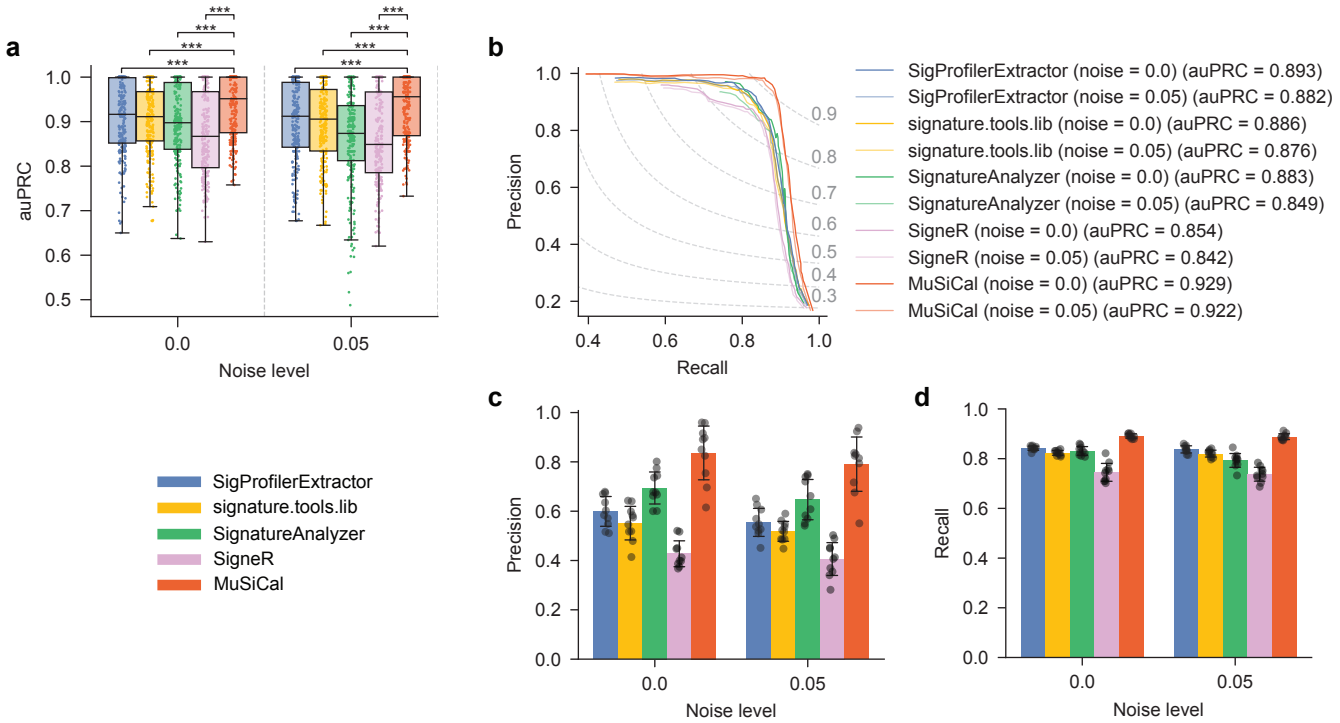
Extended Data Fig. 1 mvNMF improves the accuracy of *de novo* signature discovery in simulated datasets with SBS signatures. Synthetic datasets are simulated from tumor type-specific SBS signatures and Dirichlet-distributed exposures for 32 tumor types with 20 replicates. Each dataset contains 200 samples 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. $n = 20$ independent simulations for each box plot. **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.



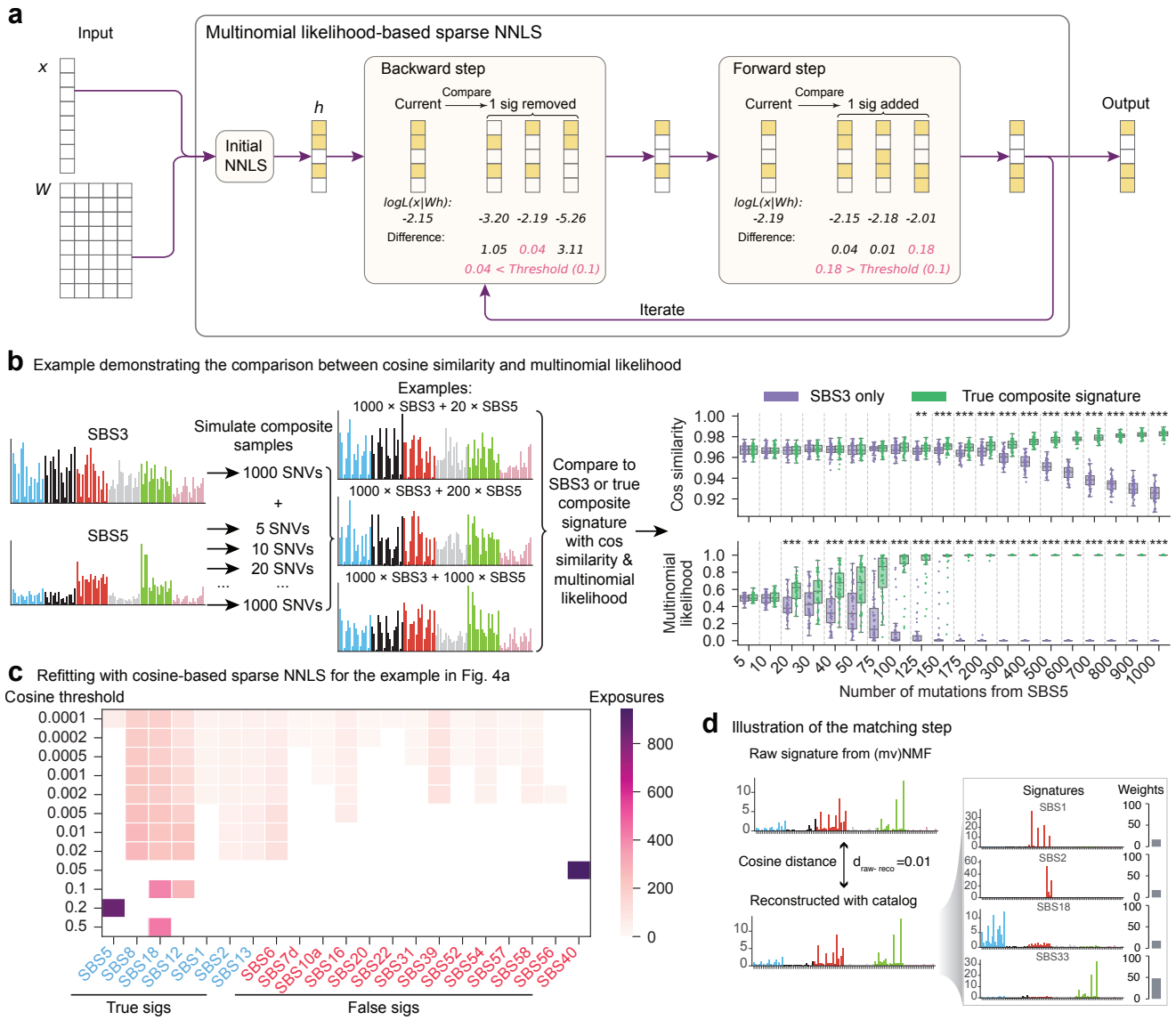
Extended Data Fig. 2 An example illustrating the method for selecting the number of signatures r. a. Data simulation scheme. A synthetic dataset is simulated from Dirichlet-distributed exposures and 6 SBS signatures (SBS1, 2, 5, 13, 18, and 40) believed to be present in cervical cancers according to 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 [61, 62]. 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.



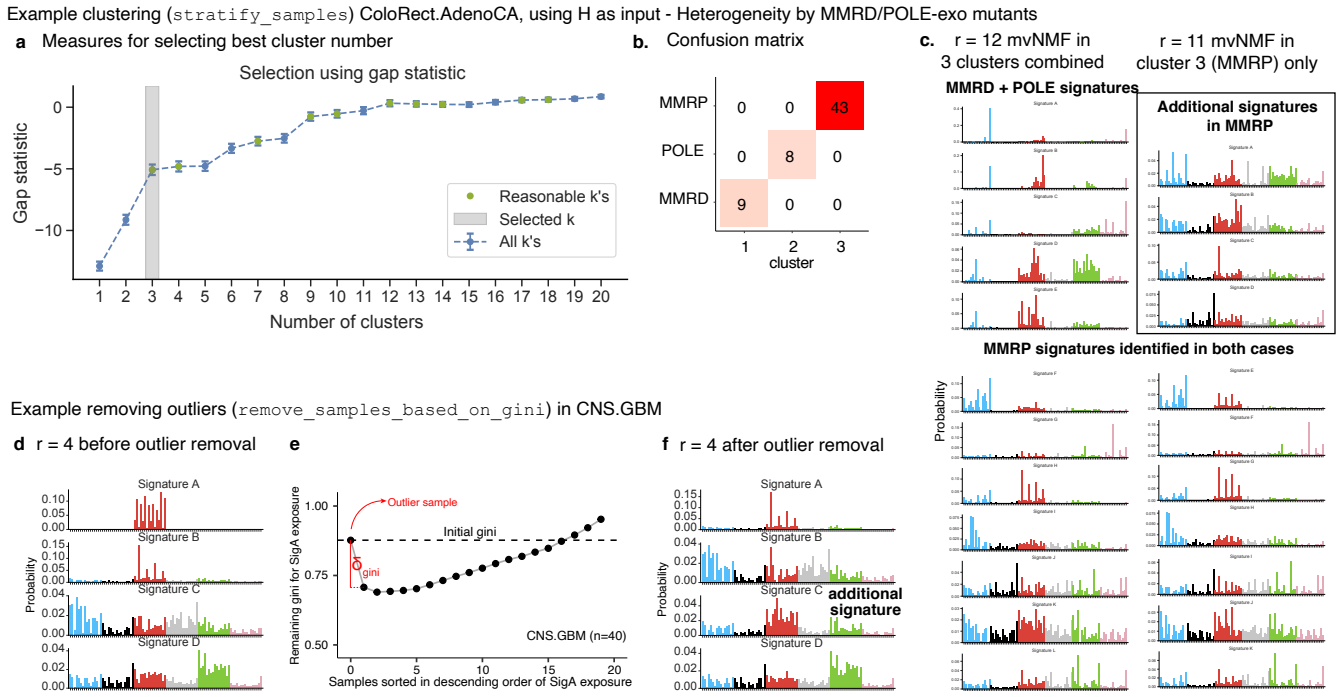
Extended Data Fig. 3 MuSiCal outperforms SigProfilerExtractor for *de novo* signature discovery at different noise levels. See Methods for details on how different levels of random noise were added. **a.** Area under precision-recall curve (auPRC) for MuSiCal and SigProfilerExtractor at different noise levels. Each box in the box plot represents 250 synthetic datasets (25 tumor types × 10 replicates). auPRC was calculated for each dataset separately, as in Fig. 3b. ***: $p < 0.0005$. p-values were calculated with two-sided paired t-tests. Raw p-values from left to right: 7.5×10^{-10}, 1.5×10^{-8}, 1.7×10^{-6}, 3.1×10^{-11}, 5.4×10^{-12}. **b.** Precision-recall curve (PRC) for MuSiCal and SigProfilerExtractor at different noise levels. Each PRC represents the average result of 250 synthetic datasets (25 tumor types × 10 replicates), as in Fig. 3c. **c.** Precision of MuSiCal and SigProfilerExtractor averaged across all tumor types at different noise levels. Recall was fixed at 0.9. Error bars indicate standard deviation over 10 replicates. **d.** Recall of MuSiCal and SigProfilerExtractor averaged across all tumor types at different noise levels. Precision was fixed at 0.98, corresponding to a false discovery rate (FDR) of 2%. The black triangle indicates the case where a precision of 0.98 was never achieved and the recall at the highest achieved precision was shown. Error bars indicate standard deviation over 10 replicates.



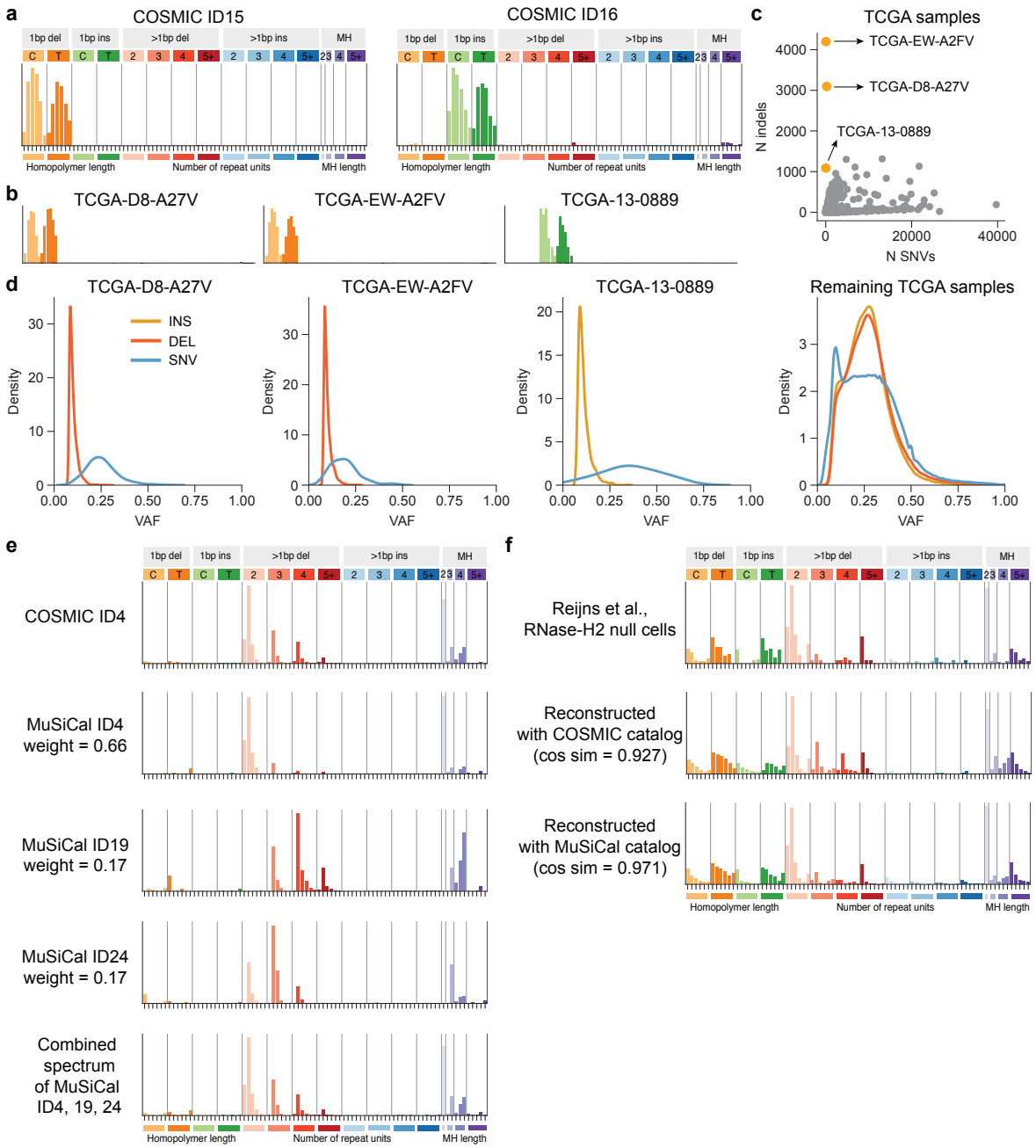
Extended Data Fig. 4 MuSiCal outperforms SigProfilerExtractor and three additional existing tools – signature.tools.lib, SignatureAnalyzer, and SigneR – for *de novo* signature discovery at different noise levels. a. Area under precision-recall curve (auPRC) for all five tools at different noise levels. Random noise were added as in Extended Data Fig. 3. Each box in the box plot represents 250 synthetic datasets (25 tumor types × 10 replicates). auPRC was calculated for each dataset separately, as in Fig. 3b. ***: $p < 0.0005$. p-values were calculated with two-sided paired t-tests. Raw p-values from top to bottom at noise level 0.0: 1.2×10^{-26}, 1.4×10^{-15}, 1.1×10^{-14}, 7.5×10^{-10}. Raw p-values from top to bottom at noise level 0.05: 1.7×10^{-25}, 1.3×10^{-20}, 5.8×10^{-16}, 3.1×10^{-11}. **b.** Precision-recall curve (PRC) for all five tools at different noise levels. Each PRC represents the average result of 250 synthetic datasets (25 tumor types × 10 replicates), as in Fig. 3c. **c.** Precision of all five tools averaged across all tumor types at different noise levels. Recall was fixed at 0.9. Error bars indicate standard deviation over 10 replicates. **d.** Recall of all five tools averaged across all tumor types at different noise levels. Precision was fixed at 0.9, corresponding to a false discovery rate (FDR) of 10%. Here, precision was fixed at a smaller value compared to Fig. 3e, Supplementary Fig. 5d, and Extended Data Fig. 3d, because a precision of 0.98 was never achieved by SignatureAnalyzer or SigneR in many cases. Error bars indicate standard deviation over 10 replicates.



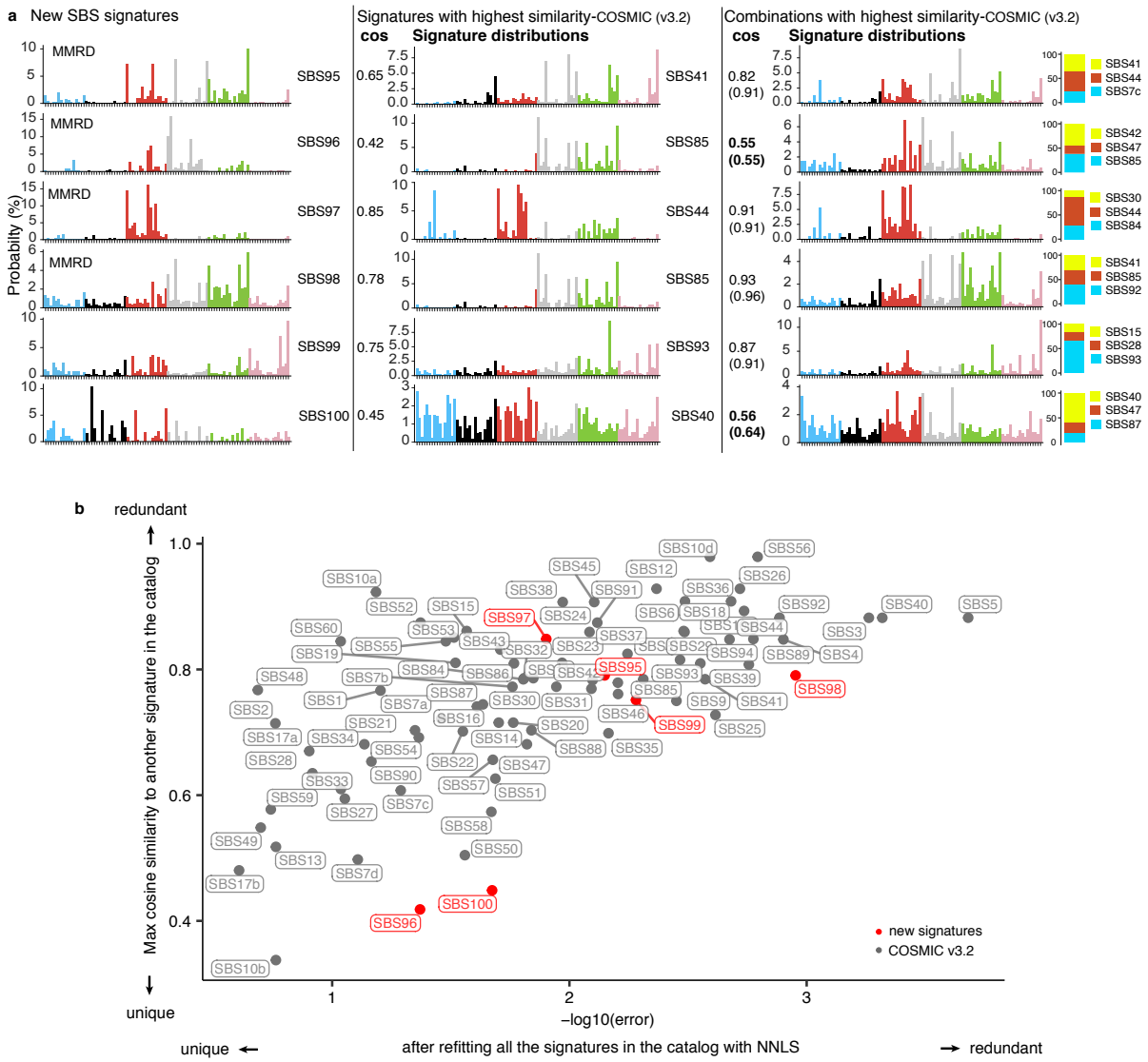
Extended Data Fig. 5 Matching and refitting with multinomial likelihood-based sparse NNLS. a. Diagram of the likelihood-based sparse NNLS algorithm. See Methods for details. **b.** An example demonstrating that multinomial likelihood is more powerful than cosine similarity for separating similar signatures. 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 two-sided 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 is 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.



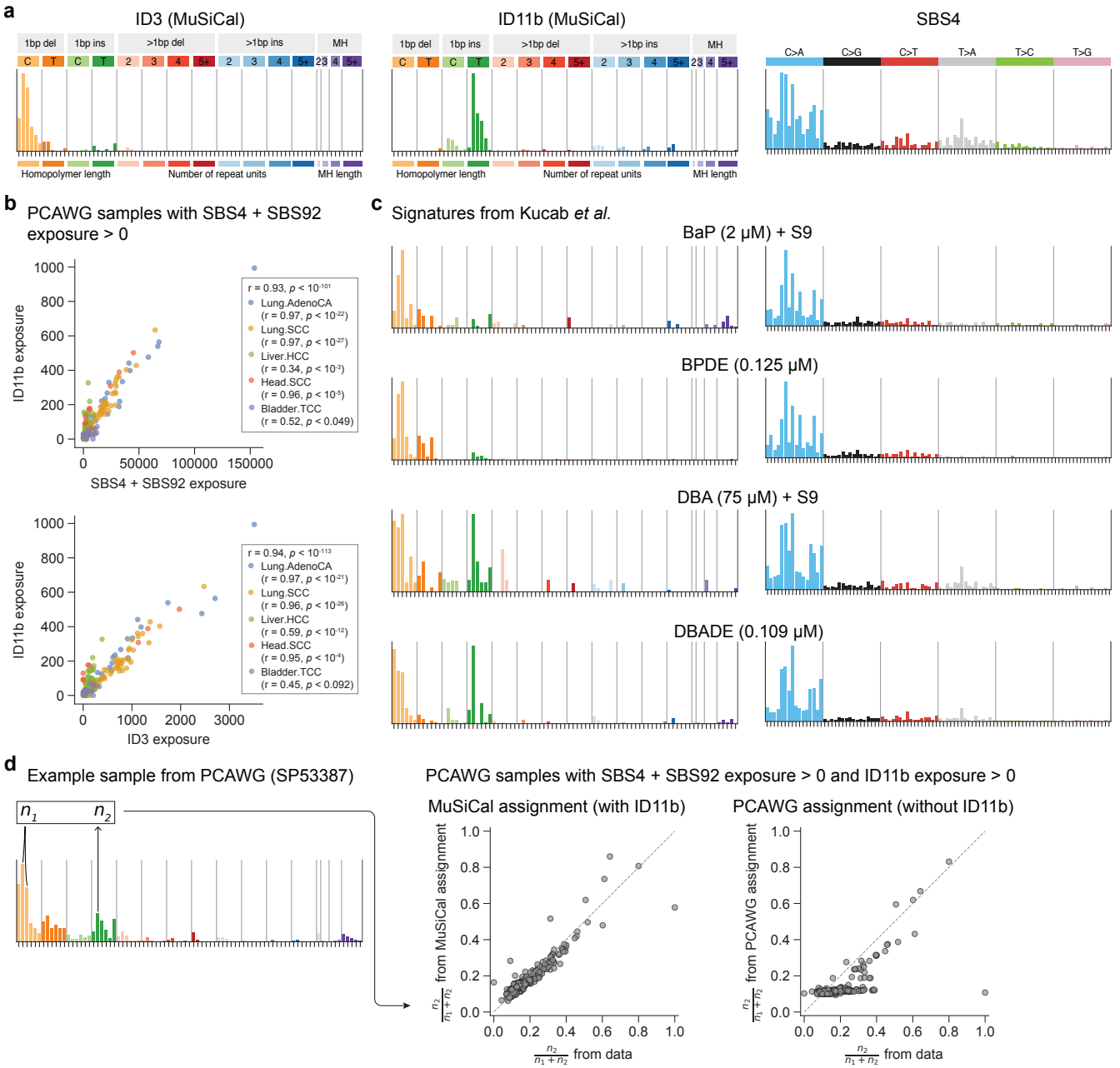
Extended Data Fig. 6 Examples illustrating cohort stratification and outlier removal with the preprocessing module. **a-c.** An example with the PCAWG ColoRect.AdenoCA dataset to demonstrate that cohort stratification can benefit *de novo* signature discovery. After an initial signature discovery with 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 (43 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. In panel (a), reasonable k means any k satisfying $\mathrm{Gap}(k) \geq \mathrm{Gap}(k+1) - s_{k+1}$, where $\mathrm{Gap}(k)$ denotes the gap statistic (indicated by dots in the plot) and s_k the standard deviation of 50 independent simulations after accounting for simulation errors (indicated by error bars in the plot). The smallest reasonable k is chosen as the optimal k. See [61, 62] for more details. **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.



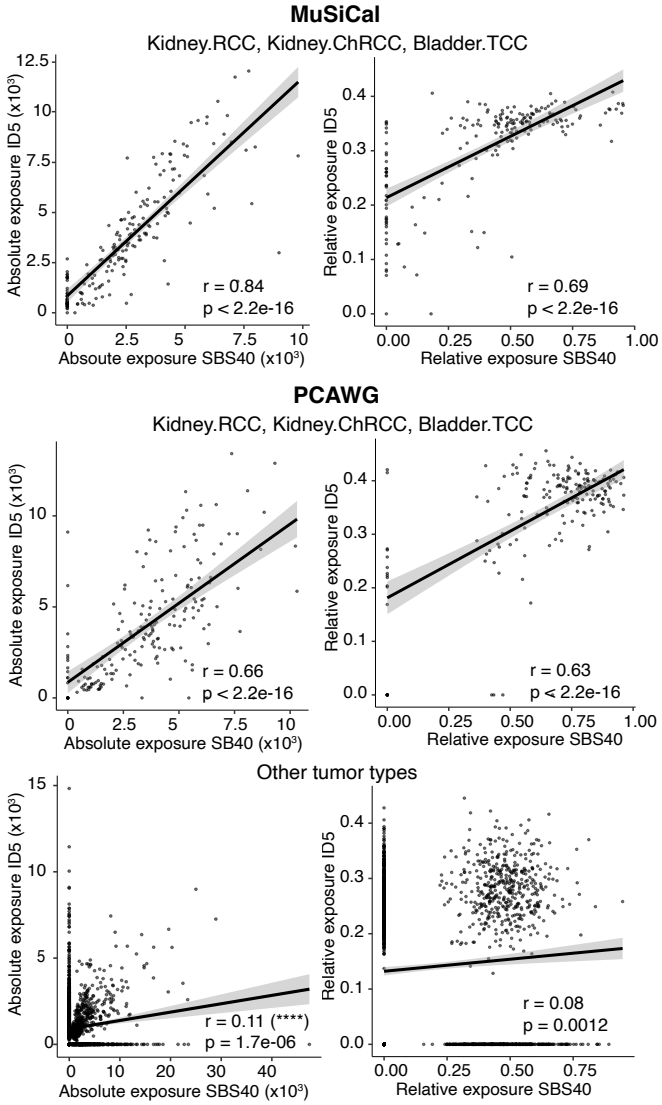
Extended Data Fig. 7 Further details related to the updated catalog of ID signatures. a. Spectra of COSMIC ID15 and 16. These two signatures are not discovered from our PCAWG reanalysis with MuSiCal. **b.** Indel spectra of the three TCGA whole-exome sequenced samples from which COSMIC ID15 and 16 are discovered by the PCAWG consortium [5]. **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.** Variant allele frequency (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 [43] (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. 8 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. 9 ID11b is potentially associated with tobacco smoking. a. Spectra of signatures known to be associated with tobacco smoking (ID3 and SBS4) and ID11b. SBS92 (not shown) is a recently discovered signature also associated with tobacco smoking [17]. **b.** ID11b exposure is correlated with both SBS4 + SBS92 and ID3 exposures. PCAWG samples related to tobacco smoking, as indicated by nonzero exposures of SBS4 + SBS92, are selected. The per-sample exposures are then plotted for ID11b vs. SBS4 + SBS92, and ID11b vs. ID3. Pearson correlation coefficients and the corresponding p-values are annotated. **c.** ID (left) and SBS (right) signatures observed in cells exposed to different chemicals present in tobacco smoke. Mutation data is obtained from [7]. 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. 10 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). Error bands (shaded areas) indicate 95% confidence intervals of the linear regressions.