StarSignDNA: Signature tracing for accurate representation of mutational processes

Christian Domilongo Bope[1,5], Sumana Kalyanasundaram[2], Knut D. Rand[2], Sigve Nakken[1,3,4], Ole Christian Lingjærde [1,2,6]*, Eivind Hovig [1,2,3]*

1 Centre for Bioinformatics, University of Oslo, Oslo, Norway
2 Department of Informatics, Faculty of Mathematics and Natural Sciences, University of Oslo, Norway
3 Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway
4 Centre for Cancer Cell Reprogramming, Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway
5 Department of Mathematics and Department of Computer Science, Faculty of Sciences, University of Kinshasa, Kinshasa, Democratic Republic of Congo
6 Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway

* corresponding ehovig@ifi.uio.no and ole@ifi.uio.no

Abstract

All cell lineages accumulate mutations over time, increasing the probability that some lineages eventually become malignant. Many of the processes responsible for generating mutations are leaving a characteristic footprint in the genome that allows their presence to be detected. However, the mutational pattern in a tumor is usually the combined result of multiple mutational processes being at work simultaneously, and the problem of disentangling the different footprints and their relative impact then becomes a deconvolution problem. Several algorithms have been developed for this purpose, most of them involving a factorization of the mutation count matrix into two nonnegative matrices, representing respectively the underlying mutational signatures and the relative weighting of (or exposure to) these signatures. Here, we introduce the StarSignDNA algorithm for mutational signature analysis, which offers efficient refitting and *de novo* mutational signature extraction. StarSignDNA is capable of deciphering well-differentiated signatures linked to known mutagenic mechanisms and suggesting clinically relevant predictions for a single patient. The package offers a command line-based interface and data visualization routines.

Keywords: Mutational signature, Mutational processes, Cancer Genomics, Nonnegative matrix factorization

Author summary

StarSignDNA is a novel algorithm for identifying mutational signatures from cancer sequencing data. It excels in low mutation count scenarios, balancing signature detection and true signature discovery. StarSignDNA improves prediction accuracy and biological validity by addressing challenges such as overfitting and underfitting, achieving optimal

variable selection and shrinkage, and ensuring interpretability. In refitting, it reduces prediction variance and handle sparsity, resulting in more precise predictions. In *de novo* extraction, it improves the detection of challenging signatures and achieves better alignment between detected and known signatures through unsupervised cross-validation. Its unique features, including prediction confidence and customizable reference signatures, offer valuable insights for clinical single-sample analysis.

Introduction

A diverse set of mutational processes shape cells that ultimately may develop into cancers. These processes encompass endogenous cellular mechanisms, such as errors in the DNA replication and repair machinery, as well as exogenous sources of DNA damage, like tobacco and UV exposure [1]. The imprints of mutational processes that have occurred during tumorigenesis are reflected in distinct spectra of acquired DNA aberrations in tumor samples, and these aberrations are now effectively captured with high-throughput sequencing technologies. A characteristic pattern of mutations caused by a mutational process is now commonly referred to as a mutational signature [2]. Accurate identification of mutational signatures across tumor samples is not only critical for our understanding of the mechanisms driving tumor development, but it can also add novel value when it comes to patient stratification toward personalized treatment regimes [3]. While distinct patterns caused by somatic copy number aberrations and insertion/deletion type of mutations have been found [4], we here focus on the signatures exerted by single base substitutions (SBS), the most common type of aberration in cancer genomes.

Several algorithms and computational tools have been developed for detection and estimation of mutational signatures in cancer genomes. These methods can broadly be classified into two main categories: discovery of novel signatures in a collection of sequenced tumor samples (also referred to as *de novo* signature identification), and fitting known reference signatures to an existing mutational profile (signature reconstruction or re-fitting). Non-negative Matrix Factorization (NMF), or its Bayesian version (BNMF) [5–7], [1] [8–13], is a matrix decomposition approach that breaks down a complex matrix into smaller, more manageable subproblems. Apart from NMF-based approaches, there are also probabilistic methods such as *Emu*, which uses Expectation Maximization. In *Emu*, the underlying probabilistic model assumes that input samples are independent, and the number of mutational signatures is inferred using the Bayesian Information Criterion (BIC) [14]. Another method, *pmsignature*, employs a flexible approach that simultaneously reduces the number of estimated parameters and allows modification of key contextual parameters, such as the number of flanking bases [15]. The collective effort of identifying signatures is collated in the Catalogue Of Somatic Mutations In Cancer (COSMIC, `https://cancer.sanger.ac.uk`), which also links each signature to the proposed underlying processes. The refitting algorithms aim to search for the best combination of established COSMIC signatures that explains the observed mutational catalogue. They achieve this by mapping the catalogue into the multidimensional space of all non-negative linear combinations of the N-established signatures.

However, despite advances in the field, challenges remain in accurately characterizing mutational signatures. A key challenge is to recover the signature composition robustly when the mutation data is limited, which is a common scenario for many samples sequenced through whole-exome or targeted assays. A low mutation count can lead to unreliable discovery of some active signatures, increased prediction variance and higher fitting error [16]. To address this challenge, a common practice in signature discovery is the use of regularisation. Regularisation promotes sparsity in the model and essentially

encourages fewer non-zero coefficients. This can help to prevent overfitting, but it can also introduce a bias into the predictions which can be negligible when using a better optimization approach. Several methods achieve regularisation in different ways; for example by introducing constraints on the model [17], by using a Bayesian variant of Non-Negative Matrix Factorization (NMF) with a penalty on model complexity [5], by employing an L1-based regularisation for sparsity on the signatures [12] or by leveraging the L1 norm penalty to promote sparsity [18]. These techniques aim to achieve better predictions by reducing the variance of the estimates.

In addition to the aforementioned challenges, current *de novo* signature discovery methods face the following weaknesses: (i) Determining the optimal number of signatures (K) to extract: There is currently no definitive method for selecting the ideal number of signatures (K) to extract from the data. (ii) Balancing the number of signatures and cosine similarity: Achieving a good balance between the number of extracted signatures (K) and their similarities to known signatures (measured by cosine similarity, typically with a default threshold of 0.8) is desirable. It is important to remember that simply fitting the data well (minimizing residual error) doesn't necessarily reflect actual biological processes [11] [19]. The common practices to determine the number of signatures are: (1) The number of signatures (K) should be selected in such a way that higher K will imply lower residual error [11]; (2) K is selected based on the residual error and maximizing the reproducibility of signatures; (3) Calling signatures hierarchically on subsets of samples, adding more signatures to fit every sample. The first two approaches are arbitrary and the third leads to overfitting, since there is a need to use more signatures for better fitting without any constraint. Recently, different methods have been developed to improve the selection of the number of signatures. SignatureAnalyzer uses automatic relevance determination, which starts with a high number of signatures and attempts to eliminate signatures of low relevance [5]. SparseSignatures both identifies the signatures in a dataset of point mutations, calculates their exposure values, and tends to infer signatures with spiking profiles [12] and add background noise. SUITOR is an unsupervised cross-validation method that selects the optimal number of signatures to attain the minimal prediction error in the validation set and extends the probabilistic model to allow missing data in the training set [20].

In this paper, we introduce StarSignDNA, an NMF model that offers *de novo* mutation signature extraction, as well as refitting to existing mutational signatures. The algorithm combines the use of regularisation to allow stable estimates with low sample sizes with the use of a Poisson model for the data to accommodate low mutational counts. StarSignDNA overcomes the aforementioned limitations for both refitting and *de novo*. StarSignDNA refitting is p[timized in the following ways:(1) utilizes LASSO regularization to minimize the spread (variance) in exposure estimates. This helps ensure more consistent and reliable exposure values, reducing the overall variability. (2) Stable exposure estimation: ensuring a minimum number of accurate processes with low fitting errors to describe the mutational catalogue of a cancer genome. (3) Improved F1 score: Achieves a higher F1 score by reducing both false positives (overfitting of signatures) and false negatives. (4) Single-Patient Suitability: Provides confidence levels on the predicted processes, making it suitable for single-patient evaluation of mutational signatures. (5) Weight matrix integration: Incorporation of the distribution of triplets in a reference genome/exome or normal tissue from the same patient (weight matrix). The StarSignDNA *de novo* module is optimized in the following ways: (1) Optimal model selection: It combines unsupervised cross-validation and the probability mass function as a loss function to select the best combination of the number of signatures (K) and regularisation parameters (λ). (2) Unbiased signature detection: The algorithm avoids introducing bias towards unknown

signatures and achieves a better balance between the number of signatures (K) and the cosine similarity threshold (above 0.8). (3) High signature sensitivity: Demonstrates high sensitivity in detecting signatures. The package is available at `https://github.com/uio-bmi/StarSignDNA` and can be installed through PyPI.

Methods and materials

Mutational signatures

Single base mutations can be classified in 96 categories $X[Y \to Z]W$, where Y denotes the reference nucleotide, Z is the mutated variant, and X and W represent the two flanking nucleotides. Let $\mathbf{m} = (m_1, \ldots, m_{96}) \in \mathbb{R}^{96}$ denote the number of mutations in each category, with the convention that the categories are lexicographically sorted on the mutation names. Suppose this mutation pattern is the combined effect of several mutational processes with distinct mutational signatures $\mathbf{s}_i$, $i = 1, \ldots, K$. Then the observed mutational spectrum $\mathbf{m}$ may be modeled as a random sample from a distribution $p(\mathbf{m}|\boldsymbol{\theta})$ with mean $\boldsymbol{\mu} = e_1\mathbf{s}_1 + \cdots + e_K\mathbf{s}_K$. where all the weights are nonnegative and some may be equal to zero. Both the signatures and the weights $e_1, \ldots, e_K$ may require estimation from a collection of observed mutational patterns (*de novo* signature extraction). If the signatures are known, we only have to estimate the weights (signature refitting). The proposed method can perform both operations.

Distinct mutagenic processes are assumed to leave different footprints in the genome in the form of different mutational signatures. Each mutational signature is a numerical vector $\mathbf{s}_k = (s_{1,k}, \ldots, s_{96,k})^T \in \mathbb{R}^{96}$, specifying the relative frequency of occurrence of each of the 96 mutation types. With no loss of generality, we may assume that the components of a signature sum to 1 (this implicitly requires that at least one mutation type is present in a signature, which is a reasonable assumption). Given a collection of K mutational signatures we define the signature matrix as $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_K)^T \in \mathbb{R}^{K\times 96}$.

Suppose we have an $N \times 96$ matrix $\mathbf{M} = (m_{ij})$ of observed mutation counts for N samples, and that we also have a $K \times 96$ signature matrix $\mathbf{S}$. We wish to determine the exposure matrix $E = (e_{ik}) \in \mathbb{R}^{N\times K}$, where each row contains the signature weights for a particular sample, with the constraint that all the elements of E should be nonnegative. We assume that the mutation counts follow a normal distribution $p(\mathbf{m}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{m}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with a sample-specific mean $\boldsymbol{\mu}$ and the variance-covariance matrix $\boldsymbol{\Sigma} = \sigma^2 I$, where I is the identity matrix. The exposure matrix can then be found by minimizing

$$L(E) = ||M - ES||_F^2. \tag{1}$$

where $||A||_F^2 = \sum_{ij} a_{ij}^2$ denotes the squared Frobenius norm of A. In practice, the normality assumption depends on the counts being sufficiently separated from zero. Mutational data obtained from low-coverage sequencing may not satisfy this assumption, leading to a biased estimate for E. To circumvent this problem, some methods allow different distributional assumptions about the data. For example, *sigfit* [21] assumes that the observed mutation counts follow either a Gaussian, a Poisson, or a negative binomial distribution, while Li et al., [18] assumes that the mutations follow a multinomial distribution. The quality of *de novo* mutational signatures derived by NMF is also dependent on the number of samples available for analysis and the degree of consistency between their mutational spectra. For small sample sizes, the use of regularization can significantly improve the stability of the estimated signatures [12]. An early example of regularization in NMF is the method Sparse Nonnegative Matrix Factorization *SNMF/L* [22]. Here, the loss function is extended to include two penalty terms:

$$L(E) = ||M - ES||_F^2 + \lambda_1||E||_1 + \lambda_2||S||_2^2 \tag{2}$$

The first penalty term imposes a soft constraint on the L^1-norm of the exposure matrix E, while the second penalty term imposes a soft constraint on the L^2-norm of the signature matrix S. The effect of these constraints is to force the solution to balance a good fit to the data (the first term) with a small norm of E and S. The two different norms applied have different effects on the estimates; the L^1-norm will tend to force several components of E to be zero, while the L^2-norm will tend to force components of S to be smaller, albeit not zero. Thus, the model will produce an NMF decomposition that is sparse in E and not in S. The authors also describe a variant of this method called *SNMF/R* in which an L^1-norm is applied to S and an L^2-norm is applied to E, resulting in an NMF decomposition that is sparse in S and not in E. The hyperparameters $\lambda_1, \lambda_2 > 0$ control the balance between a good fit to the data (the first term) and a high penalty on E and S, respectively. A more recent example of regularization in NMF is *sigLASSO* [18]. This is a refitting algorithm and the loss function is

$$L(E) = l(E|M) + \lambda \sum_{i=1}^{N} \sum_{k=1}^{96} |e_{ik}| \tag{3}$$

where $l(E|M)$ is the probability of observing M conditional on the true exposure levels being E. An even more recent example of regularization in NMF is *SparseSignatures* [12], where the original loss function is replaced by

$$L(E, E_0) = ||M - ES - E_0 S_0||_F^2 + \lambda ||S||_1 \tag{4}$$

Here, the reconstruction of M is split into two components, ES and $E_0 S_0$, where the first component is subject to regularization and the second is not. The regularization of the first component is imposed by the addition of a penalty term $\lambda ||S||_1 = \lambda \sum |s_{ij}|$ that forces the L^1-norm of the signature matrix S to be small. The hyperparameter λ controls the balance between a good fit to the data (the first term of the loss function) and a small L_1-norm of the signature matrix. Note that the *sparse signatures* algorithm will produce an NMF decomposition that is sparse in S and not in E, hence bears resemblance to the previously described SNMF/R algorithm.

The StarSignDNA algorithm

Here, we propose an NMF model that combines the use of regularization to allow stable estimates with low sample sizes and the use of a Poisson model for the data to accommodate low mutational counts. Specifically, we assume that the counts $\mathbf{m} = (m_1, \ldots, m_{96})$ for a given sample are drawn from independent distributions

$$P(m_j) = \frac{\mu_j^{m_j} e^{-\mu_j}}{m_j!} \tag{5}$$

where $\mu_j = e_1 \mathbf{s}_{1j} + \cdots + e_K \mathbf{s}_{Kj}$. Assuming N independent observations, the log-likelihood with an L_1 penalty on the exposure levels becomes

$$L(\mathbf{e}) = -\sum_{j=1}^{96} \left[m_j \log \left(w_j (\mathbf{S}\mathbf{e})_j \right) - w_j (ES)_{ij} \right] + \lambda ||\mathbf{e}||_1) \tag{6}$$

where $\lambda > 0$ is a penalty parameter and w_{ij} are known weights that allow the estimate to be adjusted for the prior probability of observing mutation category j in the ith sample.

The optimization is performed in two steps. In the first step, we compute the localized gradient $h_r(e)$ of the unpenalized likelihood respective to the mutational

exposures levels e_{ir} for each sample and a new constant j that denotes the sample i at the given time

$$h_r(\boldsymbol{e} \mid \boldsymbol{m}) = \sum_{k}^{96} \left(\frac{m_k s_{rk}}{\sum_q e_q s_{qk}} - w_k s_{rk} \right) \tag{7}$$

In the second step, the global gradient, denoted $g_i(\boldsymbol{e})$, is computed. The global gradient is the direction of steep descent, discontinuous at every point at which the penalized log-likelihood is not differentiable.

$$g(\boldsymbol{e})_r = \begin{cases} h_r(\boldsymbol{e} \mid \boldsymbol{m}) - \lambda \quad sign(e_r), & \text{if } e_r > 0 \\ h_r(\boldsymbol{e} \mid \boldsymbol{m}) - \lambda \quad sign(h_i \beta) & \text{if } e_r = 0 \quad and \quad |h_r(\boldsymbol{e} \mid \boldsymbol{m})| > \lambda \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where

$$sign(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x = 0 \\ -1, & \text{if } x < 0 \end{cases}$$

StarSignDNA Refitting Algorithm

Input: Mutational count vector $\mathbf{m} = (m_1, \ldots, m_{96}) \in \mathbb{R}^{96}$and weight vector $\mathbf{w} = (w_1, \ldots, w_{96}) \in \mathbb{R}^{96}$ for a single sample, and a signature matrix $\mathbf{S} = (s_{ik}) \in \mathbb{R}^{R,96}$.

Step 1: Initialize parameters: $\boldsymbol{e}^0 = \boldsymbol{e}_0$
Step 2: For $i = 0, 1, 2, \cdots$

1. Compute the local and global gradients using equations 7 and 8.
2. Compute step sizes t_{opt} and t_{edge}. T_{opt} is computed using a Taylor approximation, constrained to ensure differentiability of the log-likelihood (equation 6) as follow:

$$t_{opt} = -\frac{l^{'}(e; g_i(e))}{l^{''}(e; g_i(e))}$$

   T_{edge} is the upper bound of the step size, ensuring the log-likelihood remains within the defined domain and computed as follow:

$$t_{edge} = \min_i \{ -\frac{e_i}{g_i(e)} : sign(e_i) = -sign\{g_i(e)\} \neq 0 \}$$

3. **If** $t_{\text{opt}} \geq t_{\text{edge}}$:

$$e^{(i+1)} = e^{(i)} + \min\{t_{\text{opt}}, t_{\text{edge}}\} g(e^{(i)})$$

4. **Else If** $t_{\text{opt}} < t_{\text{edge}}$:
   (a) Switch to the NR algorithm.
   (b) Update using NR:

$$e^{i+1} = \begin{cases} e^{(i)} + t_{\text{edge}} g(e^{(i)}), & \text{if } t_{\text{opt}} \geq t_{\text{edge}} \\ e_{\text{NR}}^{(i+1)}, & \text{if } t_{\text{opt}} < t_{\text{edge}} \text{ and } \text{sign}(e_{\text{NR}}^{(i+1)}) = \text{sign}(e_{+}^{(i)}) \\ e^{(i)} + t_{\text{opt}} g(e^{(i)}), & \text{otherwise} \end{cases}$$

Step 3: Convergence Check: If $\left|\frac{\hat{e}-e}{e}\right| < \epsilon$, then break.

Step 4: Return $e^{(i+1)}$.

StarSignDNA *de novo* Algorithm

Step 1 Idem (unchanged from previous steps).

Step 2: Run *a snakemake* grid search to select the optimal value of K and λ

Step 3: Initialize e^0 and S^0.

Step 4: Choose the range of values to test for K (number of signatures) and λ (level of sparsity).

Step 5: For each pair of parameter values (K and λ), perform cross-validation.

Step 6: Update e and S:

1. Update e^{i+1} by maximizing:

$$\arg\max_{e} L(M, e, S, O) \quad \text{with respect to } e$$

2. Update S^{i+1} by maximizing:

$$\arg\max_{S} L(M^T, e^T, S^T, O^T)^T \quad \text{with respect to } S$$

Step 7: Compute the mean squared error (MSE) and the probability mass function (PMF) for each (k, λ) pair.

Step 8: Choose the values of K and λ that correspond to the lowest across all cross-validation.

Step 9: Using the selected values for K and λ, run the StarSignDNA *de novo* module on the complete matrix M. This generates the final values of S and e.

Materials

Simulated data

We first evaluated the performance of StarSignDNA using a simulation dataset. For a fair and consistent comparison of different methods, it is crucial to have true underlying processes as a reference and then evaluate the ability of the different methods to detect them. Four categories of simulated datasets were considered. The first three categories were used to validate the refitting algorithm, and the last category was used to assess the *de novo* algorithm. Datasets of the first three were generated using N mutation vectors, where N is the count of mutation in a particular three-mer context, by drawing the *th* component of the *th* mutation vector from the non-negative binomial distribution with mean and variance of 10.1.

The first category consisted of 16 datasets (Table 1), which incorporated variations in exposure weights and the number of mutations (K=100, 500, 1000, and 1000), where K is the number of mutations per sample to analyze the sensitivity and precision.

In the second (Table 2) and third categories (Table 3), we considered two subsets of the COSMIC signatures with different characteristics (see Supplementary Figure 1). The first subset consisted of signatures that were close to all other signatures (hence hard to uniquely identify) and the second subset consisted of signatures that were distinct from all other signatures (hence easy to identify). Technically, the first subset was found by identifying the five signatures with the highest average cosine similarity to all other signatures and the highest cosine similarity to at least one other signature, while the second subset was found by identifying the five signatures with the lowest average and maximum cosine similarity to the other signatures. The first subset consists

of the signatures SBS3, SBS5, SBS40c, SBS92, and SBS94, which all have broad support (many non-zero components), and such signatures are commonly referred to as flat signature. The second subset consists of the signatures SBS7d, SBS10b, SBS13, SBS17b, and SBS28, which all have narrow support (few non-zero components), and such signatures are frequently referred to as common since they are more easily identifiable.

To evaluate and compare the performance of *de novo* signature detection, the fourth category of simulated data was used. For this, we considered N = 146 prostate samples, subject to whole genome sequencing (WGS) within the PCAWG. We then selected a set of four signatures, namely SBS1, SBS5, SBS18, and SBS40, from the COSMIC signatures known to be prevalent in prostate cancer [5]. DeconstructSigs was used to fit these selected signatures to the data and generate a reconstructed mutational catalog count. To ensure consistent comparison, 50 replicates were considered [16].

Mutational signatures in patient-derived exomes and genomes

In total, 8,893 whole exome sequencing (WES) samples across 33 cancer types from The Cancer Genome Atlas (TCGA) (`https://portal.gdc.cancer.gov/`) were used to evaluate the refitting method. A single TCGA breast cancer sample was used to demonstrate StarSignDNA's ability to incorporate confidence into predictions using the 2.5% percentile to the 97.5% percentile of the bootstrapped exposure levels.

De novo algorithm was evaluated using 3,308 WGS data selected from The Pan-Cancer Analysis of Whole Genomes (PCAWG) and the Hartwig Medical Foundation (Table 5).

Results

Refitting Mutational signatures

StarSignDNA validation on simulated data

Using the first category dataset (Table 1), which varied exposure weights and the number of mutations (K), we evaluated the precision and sensitivity of six mutational signature algorithms: StarSignDNA, deconstructSigs, MutationalPatterns, sigLASSO, SigProfilerSingleSample (SigProfilerSS), and sigfit.

The weights assigned to the four signatures represent the ground truth. To achieve a more comprehensive comparison, we employed a wider range of weights encompassing scenarios with high sensitivity and subtle signature profile deviations. Furthermore, we investigated the impact of the number of mutations per sample on the algorithms' performance by considering scenarios with 100, 500, 1000, and 10000 mutations.

The result shows (Fig 1) that all the methods perform well, except (Fig 1C-1D) where SigProfilerSS and sigLASSO show some deviation from the ground truth. These results indicate that StarSignDNA, MutationalPatterns, sigfit, and deconstructSigs outperform SigProfilerSS and sigLASSO when more than two signatures are considered. When considering the higher number of mutations (K = 500, 1000, 1000), the performance is improved for SigProfilerSS and sigLASSO indicate the challenge of both methods when considering a low number of mutations per sample (around 100).

Common signatures are a group of signatures from COSMIC with low similarity to others, appearing isolated as shown in (Supporting Figure 1). This isolation makes it easier for most prediction methods to detect them with high precision, sensitivity, and low fitting error. To assess the performance of refitting algorithms on common signatures, we considered five specific signatures: SBS7d, SBS10b, SBS13, SBS17b, and SBS28. We employed the second category of simulated data for this analysis (Table 2).

Compared to the results (Fig 1C and 1D), SigProfilerSS exhibits a higher deviation from the ground truth when considering more signatures. It also struggles to infer signatures SBS17b and SBS28. Further analysis as depicted (Fig 2D) illustrates that SigProfilerSS and sigLASSO have a higher fitting error of 0.4, while StarSignDNA, deconstructSigs, MutationalPatterns, and sigfit achieved low fitting error of 0.1, and stable prediction (S1 Table, S2 Table, and S3 Table). The fitting error was computed as follows:

$$Error = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{5}(\widehat{E_{ki}} - E_{ki}^{True})^2$$

Where N is the number of samples, k is the number of reference signatures, $\widehat{E_{ki}}$ is the estimated exposure, and E_{ki}^{True} is the ground truth exposure.

The dimension of reference signatures significantly impacts the performance of refitting algorithms [23, 24]. StarSignDNA, MutationalPatterns, deconstructSigs, and sigfit outperform SigProfilerSS and sigLASSO when using a limited number of reference signatures (four in Fig 1 and five in Fig 2). A reliable refitting algorithm should ideally maintain good performance regardless of whether a reduced or a larger set of reference signatures is used.

To assess the stability of the six methods when the reference signature set changes, we evaluated their performance using the full COSMIC v3.4 reference, containing 86 signatures. Increasing the number of signatures leads to higher uncertainty and overfitting. The F1 score metric was employed to quantify this effect.

$$F_1 \quad score = 2 \times \frac{Precision \quad \times \quad Sensitivity}{Precision \quad + \quad Sensitivity}$$

This F1 score takes both precision and sensitivity into account, providing a more comprehensive measure of the performance of a fitting algorithm.

StarSignDNA performs best for simulations with a low number of mutations (Fig 3,S4 Table, S5 Table, and S6 Table). It achieves the highest F1 score (0.98), followed by MutationalPatterns (0.96), deconstructSigs (0.95), SigProfilerSS (0.90), sigLASSO (0.77), and sigfit (0.59). For simulations with around 500 mutations, StarSignDNA, MutationalPatterns, and deconstructSigs maintain a high F1 score of one. SigProfilerSS, sigLASSO, and sigfit also perform well with an F1 score of 0.99. Interestingly, all six methods achieve a perfect F1 score (one) for simulations with around 1000 mutations. These results suggest that StarSignDNA is the recommended tool for low mutation scenarios, due to its superior performance.

The COSMIC reference profile includes both common and challenging flat signatures, such as SBS3 (related to DNA damage repair) and SBS5 (present in virtually all samples). These flat signatures exhibit higher similarity to others and appear clustered together (as shown in Supporting Figure 1), making them harder to detect. To evaluate StarSignDNA's performance on these difficult-to-detect flat signatures, the third category of simulations (Table 3) was used. This category explores the impact of exposure values and mutation counts on the performance of refitting algorithms, highlighting the challenges posed by flat signatures.

The analysis began with only five COSMIC v3.4 reference signatures (SBS3, SBS5, SBS40c, SBS92, and SBS94). StarSignDNA, sigfit, MutationalPatterns, and deconstructSigs performed better at inferring flat signatures as compared to SigProfilerSS and sigLASSO, which deviated from the ground truth (Fig 4A, 4B, 4C, S7 Table, S8 Tableand S9 Table). First four methods had the lowest fitting errors, while SigProfilerSS and sigLASSO had higher fitting errors of 0.1 (Fig4 D). Further analysis using the full set of 86 COSMIC v3.4 signatures on simulations with 100, 500, and 1000 mutations per sample (SimF(100), SimF(500), and SimF(1000)) revealed that for 100

mutations, StarSignDNA achieved the highest F1 score (0.28). SigProfilerSS followed closely at 100 mutations (F1 score: 0.26), whereas deconstructSigs and sigLASSO had lower scores (around 0.13). MutationalPatterns and sigfit performed poorly in this scenario, with F1 scores of 0.09 and 0, respectively.

As the mutation count increased to 500 per sample, StarSignDNA outperformed deconstructSigs (F1 score: 0.45 vs. 0.4). sigLASSO and SigProfilerSS also improved, while MutationalPatterns had a moderate score (0.31). sigfit performed poorly again, with a score of 0. With 1000 mutations per sample, deconstructSigs lead (F1 score: 0.54), followed by StarSignDNA (F1 score: 0.5), while sigLASSO, and MutationalPatterns had scores around 0.4-0.42. SigProfilerSS remained competitive (F1 score: 0.4), and sigfit continued to score 0. Table 4 compares the methods using the median Pearson's correlation coefficient, showing StarSignDNA as having the highest median per-patient correlation across all datasets, followed by SigProfilerSS (Fig 5 and Table 4 (S10 Table, S11 Table, and S12 Table)), indicate that StarSignDNA excels in simulations with lower mutation counts (around 100 and 500), making it suitable for refitting analyses of both common and flat signatures. It is important to note that the level of prediction uncertainty is directly related to the number of mutations per sample. In simpler terms, fewer mutations lead to a higher degree of uncertainty in the sampling process. Fortunately, this uncertainty becomes statistically insignificant when there are many mutations per sample.

We investigated the performance of refitting algorithms for flat signatures using downsampled datasets named SimF50, SimF70, SimF80, SimF90, and SimF100 (Table 3) to assess prediction stability and its correlation with the number of mutations per sample. It aimed to identify the acceptable number of mutations needed for accurate predictions using the full COSMIC v3.4 reference signature set. Results show that StarSignDNA consistently outperformed other methods across these datasets. The highest F1 score for StarSignDNA on SimF50 was 0.22. While this is low for high-precision applications, it indicates that around 50 mutations per sample are the threshold for reliable predictions. Below this threshold, predictions become more random and uncertain. The dimension of the reference signature matrix is crucial for accurate detection with low mutation counts, and prior knowledge of the cancer type is important for selecting relevant signatures. Notably, StarSignDNA's superior performance, with higher F1 scores as compared to other methods (Fig 6), especially when limited prior knowledge is available and the full COSMIC reference set is used (S10 Table and S13 Table). By profiling StarSignDNA, deconstructSigs, MutationalPatterns, sigLASSO, and Sigfit (Supporting Figure 2), we observed that the total number of reference signatures did not significantly impact computational time. Below 40 signatures, all methods except Sigfit demonstrate fast running times. Notably, with over 80 signatures, StarSignDNA, deconstructSigs, MutationalPatterns, and sigLASSO exhibit nearly identical performance. The hyperparameter λ which controls the balance between a good fit to the data is obtained using cross-validation (Supporting Figure 3) from both WES and WGS.

StarSignDNA validation on real cohort datasets

We applied StarSignDNA to analyze 8,893 TCGA tumors across 33 cancer types (Fig 7). The result for sigLASSO and deconstructSigs can be found in (Supporting Figure 4 and Supporting Figure 5). We used the full COSMIC v3.4 signatures as the reference. The results demonstrate StarSignDNA's ability to generate sparse signature assignments which can improve prediction owing to lower estimator variance, particularly when dealing with low-count WES datasets.

StarSignDNA validation on single real data samples

In a clinical setting, analyzing a single sample is crucial. StarSignDNA addresses this by incorporating a confidence level for its predictions using bootstrapping. Here the illustration is done for a single breast cancer patient sample of TCGA-BRCA (Fig 8).

De novo Mutational signatures

StarSignDNA Validation on simulation datasets

StarSignDNA was benchmarked versus other *de novo* algorithms, including SigProfiler, an NMF-based method, SparseSignatures, a cross-validation method to select both the number of signatures and the shrinkage parameter simultaneously [12], SUITOR, a probabilistic model [20], and SigneR, a Bayesian approach [8]. The benchmarking was performed using the fourth category of simulated data sets. We generated 50 replicates datasets of 146 patients each, with a set of ground truth mutational signatures being COSMIC SBS1, SBS5, BSB18, and SBS40 (see Methods). In the *de novo* signature extraction method, identifying the optimal number of signatures (K) is critical. This value (K) reflects the underlying biological processes contributing to the mutations. To determine the optimal K, we analyzed a range of values from 2 to 18, allowing each algorithm to select the most fitting number. It is important to note that while the true K for the simulation dataset data is known to be four, there might be additional detectable processes present since the simulation dataset data was derived from real-world data. The analysis focused on two key aspects: (i) the number of signatures (K) extracted by each method and (ii) the number of mutational processes each method could recover. Although the cosine similarity threshold can be set to 0.75 when identifying novel signatures [25], in this paper a stringent threshold of 0.8 was set to define acceptable reconstructed signatures.

Considering the number of extracted signatures (K), all methods identified at least four (K=4), corresponding to the overweighted signatures in the data. SUITOR and SigneR consistently detected only the overweighted signatures across all 50 replicates. SparseSignatures exhibited some variability: K=6 in two replicates, K=4 in 21 replicates, and K=3 in 27 replicates. SigProfiler showed similar behavior with K=4 in 43 replicates, K=3 in three replicates, and K=2 in four replicates. Interestingly, StarSignDNA identified more than four signatures: K=6 in 23 replicates, K=5 in 23 replicates, K=4 in two replicates, and K=3 in two replicates. It's important to note that while finding more signatures generally reduces residual error and improves data fitting. The primary goal of a *de novo* algorithm goes beyond simply fitting the data. Ideally, it should also identify processes reflecting distinct underlying biological mechanisms. In terms of residual error, SigneR and SigProfiler followed by StarSignDNA achieved the lowest values shown in (Fig 9B). However, (Fig 9A) reveals that SigneR and SigProfiler fail to detect active processes beyond the overweighted ones. In contrast, StarSignDNA and SigneR perform better at reconstructing the original input signatures (Fig 9C). StarSignDNA's superior accuracy can be attributed to its optimization algorithm and regularization on exposures based on learned parameters (λ). SparseSignatures, on the other hand, utilizes regularization on the signatures itself, explaining its higher sparsity. Notably, StarSignDNA also exhibits a higher sparsity level compared to other methods (Fig 9D).

StarSignDNA detected SBS1 in 38 replicates, SBS5 in 40 replicates, SBS18 in 31 replicates, SBS40 in 48 replicates, and SBS6 in two replicates. SigneR detects SBS1 in 49 replicates, SBS5 in 49 replicates, and SBS18 in 37 replicates and 49 replicates. SUITOR detects SBS1 in 40 replicates, SBS5 in 49 replicates, SBS18 in 35 replicates, and SBS40 in 49 replicates. Sparse Signatures detect SBS1 in 22 replicates, SBS40 in 49

replicates, and SBS60 in nine replicates. Sigprofiler detected SBS1 in 49 replicates, SBS3 in three replicates, SBS5 in 42 replicates, SBS18 in 49 replicates, SBS40 in 46 replicates, and SBS16 in seven replicates. As show from (Fig 10), it clearly show that StarSignDNA outperforms the other algorithms by capturing signature SBS3 in 36 replicates compared to other methods (SigProfiler: three replicates, SUITOR: one replicate, SigneR, and SparseSignatures: 0 replicate). This is likely because SBS3 is a difficult-to-detect signature due to its high similarity to other signatures. This finding aligns with (Fig 9A), which demonstrated StarSignDNA's ability to detect less frequent and potentially more challenging processes. It's further corroborated by the refitting results in (Fig 5-6). The superior performance of StarSignDNA in detecting SBS3 might be attributed to its grid-search algorithm. This algorithm employs a combination of cross-validation and the probability mass function to effectively select the optimal number of active processes within the data.

De novo mutational signatures validation on real datasets

De novo mutational signature extraction was validated on real data from whole-genome sequencing (WGS) of tumors across five cancer types (Table 5). These cancer types include breast, skin, prostate, kidney, and pancreatic. The data consisted of 1,049 tumors from PCAWG and 2,924 tumors from Hartwig Medical Foundation (HMF).

Mutational signature on primary tumour data (PCAWG)

All five algorithms were run on the same datasets, allowing them to search for the optimal number of signatures (K) in the interval from two to 18 (Table 6). In this paper, a *higher signature* refers to a detected *de novo* signature with a cosine similarity equal or greater to 0.8. The signatures predicted by SigneR (K=18) and SparseSignatures (K=11) appear to contain more noise because only eight and one higher signature respectively. StarSignDNA (K=4) and SUITOR (K=4) show a better prediction ratio of 0.75 followed by SigProfiler (K=5). Looking into the actual signatures extracted when considering higher signatures, SigneR detected 18 signatures for skin cancer in PCAWG and only eight had higher signatures. Sigprofiler detected seven signatures, and only three had higher signatures. Sparsesignatures detected 11 signatures, and only one had a higher signature. SUITOR detected four signatures and only three had higher signatures. StarSignDNA detected four signatures and only three had higher signatures.

Unlike simulated data with known true signatures, real-world data lacks this information. However, we can evaluate if the detected signatures have a high correlation with the observed mutations. As presented in (Table 6), the median per-patient correlation coefficient (measured by Pearson's correlation) was used to assess how well each method fits the observed mutations in individual patients. This coefficient compares the observed mutation spectrum to the predicted one for each patient. StarSignDNA demonstrated the best fit to individual patient mutation counts (median correlation = 0.993), followed by SparseSignatures (0.992). This indicates that StarSignDNA and SparseSignatures captured the underlying mutational processes more effectively in these real-world cancer genomes. However, SparseSignatures has the lowest prediction ratio 0.09 follow by SigneR 0.44. In contrast, StarSignDNA (K=4) and SUITOR (K=4) achieved a better prediction ratio of 0.75, followed by SigProfiler (K=5). This suggests that StarSignDNA and SUITOR identified more relevant signatures with less noise compared to SigneR and SparseSignatures, even though they used a lower number of signatures (K).

Looking deeper into the actual signatures extracted, we observed the following: (i) SigneR detected 18 signatures, but only eight higher signatures, indicating potentially

lower quality signatures. (ii) SigProfiler detected seven signatures, with only three higher signatures. (iii) SparseSignatures detected 11 signatures, with one higher signature. (iv) SUITOR detected four signatures, with three higher signatures, and (v) StarSignDNA detected four signatures, higher signatures. Interestingly, StarSignDNA detected more higher signatures with higher median per-patient correlation, suggesting they capture biologically relevant mutational processes. Additionally, these signatures were frequently detected by the other four algorithms (Fig 11 and S19 Table).

Delving into the concept of sparsity, we see that StarSignDNA achieves a lower value (0.15) compared to SparseSignatures (0.16). In simpler terms, sparsity refers to the number of inactive elements (zeros) within a matrix. In the context of signatures, a lower sparsity indicates that more mutational signatures are present potentially contributing to the observed mutations. Interestingly, both algorithms employ regularization techniques to prevent overfitting, but they achieve this in distinct ways. SparseSignatures directly enforces sparsity by driving elements in the signature matrix towards zero, potentially eliminating weak but relevant signatures. In contrast, StarSignDNA focuses on the exposure values, which define the contribution of each signature. Shrinking these exposure values indirectly achieves sparsity while potentially retaining weak signatures within the exposure values themselves. This difference in regularization strategies might explain why StarSignDNA captures a wider range of mutational processes, leading to potentially more interpretable biological insights.

Overall, these results suggest that StarSignDNA performs well on primary WGS tunour data. It achieves high-fitting performance, identifies relevant signatures with less noise, and extracts sparser signatures that are more likely to represent distinct mutational processes.

Detection mutational signatures on metastatic

The same approach used for the primary tumor data was applied to the metastatic data to assess the prediction performance of five methods using a dataset of 2,924 metastatic samples. As shown in Table 7, SparseSignatures fit the counts of individual patients better (0.995), followed by StarSignDNA (0.994). The signatures predicted by SigneR (K=17) and SigProfiler (K=16) appear to contain more noise. StarSignDNA (K=3) and SUITOR (K=3) show a better prediction ratio of one followed by SparseSignatures with a prediction ratio of 0.67.

Looking into the actual signatures extracted when considering higher signature, SigneR detected 17 signatures and only eight higher signatures. SigProfiler detected 16 signatures and only eight higher signatures. Sparsesignatures detected three signatures, and two higher signatures. SUITOR detected three signatures and three higher signatures. StarSignDNA detected three signatures and three higher signatures. Similar to the PACWG result, most signatures found by StarSignDNA were highly similar to the higher signature, and frequently detected by the other four algorithms (Figure 12, S20 Table).

Similar to the findings in the primary tumour analysis, both StarSignDNA, and Sparsesignatures achieved high median per-patient correlation, indicating good fits to the observed mutations. Additionally, StarSignDNA and SUITOR maintained a higher prediction ratio, suggesting they identified more relevant signatures with less noise.

The result on primary tumour and metastatic data shows prediction stability StarSignDNA methods.

Evaluation of Mutational Signature Quality Using COSMIC Reference

To assess the quality of signatures identified by the five algorithms, this analysis focused on breast and skin cancer types using the COSMIC v3.4 database as the reference. For each cancer type and algorithm, the results included a signature matrix and an exposure matrix. The evaluation involved computing the similarity (Qij) between each pair of signature matrices, matching K signatures from one algorithm with another, and calculating a quality score (q) based on cosine similarity. Qij was defined as the maximum quality score across all pairwise comparisons for a specific cancer type. Higher Qij values indicate better alignment. StarSignDNA (Fig 13A and Fig 13B) achieved the highest signature quality for both breast (0.55) and skin (0.37) cancer, indicating greater similarity to the COSMIC reference signatures.

Discussion

The paper introduces StarSignDNA, a new algorithm for identifying mutational signatures from cancer sequencing data. StarSignDNA utilizes the L1-penalized estimation algorithm developed by Goeman [26], which employs directional Taylor approximations to maximize the penalized likelihood function. L1 regularization is used for its dual benefits: variable selection and shrinkage [18, 27, 28]. Variable selection is essential in mutational signature analysis to identify the most relevant signatures among many possible ones, enhancing interpretability. Shrinkage helps prevent overfitting by reducing the influence of weak signatures. Thus, StarSignDNA achieves a more interpretable and accurate identification of mutational signatures from cancer sequencing data.

To evaluate StarSignDNA's effectiveness, its performance was compared against six established refitting algorithms and five *de novo* signature extraction algorithms using both simulation datasets (2,046 samples) and real cancer data (12,201 samples). StarSignDNA excels in scenarios with low mutation counts, striking a balance between the number of signatures detected and the discovery of true mutational signatures.

The enhanced performance of StarSignDNA can improve our understanding of the key drivers of cancer initiation and the factors affecting the detection of mutational signatures and latent mutational processes. Over a decade after the Alexandrov et al., [18] publication of mutational signatures, various algorithms have emerged, producing similar but not identical results, raising questions about prediction accuracy and biological validity. A key challenge is applying insights from large-scale signature studies to datasets with few mutations and samples. Despite the suitability of decomposition methods for both large and small datasets, challenges remain, including preventing overfitting and underfitting in low mutation samples, achieving optimal variable selection and shrinkage with full COSMIC reference signatures, and ensuring the biological interpretability of predictions.

In refitting scenarios, StarSignDNA demonstrates significantly better performance, especially with low mutation counts. By applying L1 regularization, StarSignDNA reduces prediction variance and induces sparsity, resulting in more precise predictions as compared to L2 regularization. This approach leads to a better fit of observed mutations in individual patients, and lower fitting errors (Fig 1 to Fig 7). Analysis of four downsampled datasets (Fig 6) shows that the mutation count cutoff is proportional to the dimension of the COSMIC reference matrix used for fitting. Table 4 further underscores the improved fit of observed mutations in patients, offering substantial value for clinicians and researchers. Additionally, StarSignDNA provides confidence level estimates for single sample predictions, helping decision-makers understand the

range of likely true parameter values and make informed decisions about mutational processes, even in scenarios with inherent uncertainty.

De novo mutational signatures extraction presents two main challenges. First, accurately determining the optimal number of signatures (K) and second, identifying the underlying biological processes contributing to the mutations is essential. Typically, *de novo* signatures with a cosine similarity exceeding 0.8 are indicative of true processes. This work utilizes 10-fold cross-validation to select the optimal pair of (λ,K), where λ represents the regularization parameter and K signifies the number of signatures, this approach is recommended for balancing bias and variance during K selection [29, 30]. StarSignDNA exhibits several advantages compared to other *de novo* signature extraction tools. Notably, it demonstrates improved detection of challenging signatures to be detected, like SBS3 (Fig 4). StarSignDNA identified the SBS3 signature in over 36 replicates, whereas a competing SigProfiler only detected it in three replicates and the rest were unable to detect SBS3. This improvement is likely due to StarSignDNA's unsupervised cross-validation method. This method facilitates the selection of both K and the λ without overfitting the data. This combined optimization leads to two positive outcomes: first, a better ratio of detected signatures to the total number analyzed, and second, enhanced overall signature match quality by preventing overfitting and ensuring signatures better reflect the true underlying biological processes. One area for improvement in the *de novo* method is reducing the computational runtime of the grid search for selecting the optimal (λ,K).

De novo mutational signature extraction faces two main challenges: accurately determining the optimal number of signatures (K) and identifying the underlying biological processes contributing to mutations. StarSignDNA shows several advantages over other tools in detecting difficult signatures like SBS3. For example, StarSignDNA identified SBS3 in over 36 replicates, while SigProfiler detected it in only three. This improvement is due to StarSignDNA's unsupervised cross-validation method, which selects K and λ without overfitting, leading to a better ratio of detected signatures and enhanced overall signature match quality. However, reducing the computational runtime of the grid search for optimal (λ, K) selection remains an area for improvement.

StarSignDNA is a powerful tool for mutational signature detection, excelling in both refitting and *de novo* analyses. Its refitting module addresses mutation sampling uncertainty by using an optimization approach with L1-penalty and the Poisson distribution, making it suitable for low mutation counts. Taylor approximations are used to maximize the penalized likelihood function. For *de novo* analysis, StarSignDNA employs a grid search method with unsupervised cross-validation to determine the optimal number of signatures (K) and the regularization parameter. This approach uses the probability mass function as the cost function, achieving better alignment between detected signatures and those with high similarity (cosine similarity > 0.8) to known signatures. We believe the demonstrated performance and capabilities position StarSignDNA as a leading tool in both refitting and *de novo* contexts, with significant potential for contributions to genomics and oncology research.

In summary, StarSignDNA improves upon current mutational signature methods in both *de novo* discovery and refitting tasks. Its unique features, such as prediction confidence and customizable reference signatures, offer valuable insights for clinicians in single-sample analysis.

Tables

Table 1. The first category of simulated mutation datasets, consisting of 16 datasets, each containing 100 samples. These datasets were drawn from a negative binomial distribution and were used to explore the impact of two factors: exposure values and the number of mutations per sample. Each row represents a different set of exposure values, while each column represents one of the four mutational signatures. The number of mutations (K) varies within these datasets. A total of 100 samples were considered for each dataset.

Dataset	**SBS1**	**SBS2**	**SBS3**	**SBS4**	**Mutations (K)**
SimA	1	0	0	0	100, 500, 1000, 10000
SimB	$\frac{1}{2}$	$\frac{1}{2}$	0	0	100, 500, 1000, 10000
SimC	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	100, 500, 1000, 10000
SimD	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	100, 500, 1000, 10000

Table 2. The second category of simulated mutation datasets, consisting of 3 datasets, each containing 100 samples. These datasets were drawn from a negative binomial distribution and were used to explore the impact of two factors: exposure values and the number of mutations per sample. Each row represents a different set of exposure values, while each column represents one of the five mutational signatures. The number of mutations (K) varies within these datasets. A total of 100 samples were considered for each dataset.

Dataset	**SBS7d**	**SBS10b**	**SBS13**	**SBS17b**	**SBS28**	**Mutations (K)**
SimE	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	100, 500, 1000
SimF	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	100, 500, 1000
SimG	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	100, 500, 1000

Table 3. Performance using the third category of simulated mutation datasets, consisting of seven datasets, each containing 100 samples. Each row represents four different choices of exposure values, while each column represents one of the four mutational signatures. An additional layer of complexity is introduced by varying the total number of mutations within each dataset. This variation is denoted by the symbol (K). For instance, K = 50 indicates datasets where only 50% of the mutations from a baseline scenario (K = 100) are retained. Similarly, K = 70, K = 80, and K = 90 represent datasets with 70%, 80%, and 90% of the mutations remaining compared to the baseline, respectively. A total of 100 samples were considered for each dataset.

Dataset	SBS3	SBS5	SBS40c	SBS92	SBS94	Mutations (K)
SimE	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	50,70,80,90,100, 500, 1000

Table 4. Comparison of methods on simulated datasets. The table presents the median per-patient correlation measured as the median Pearson's correlation coefficient between the observed mutation spectrum and the predicted mutation spectrum for each patient.

Methods	SimF(100)	SimF(500)	SimF(1000)
StarSign	0.51	0.76	0.87
sigLASSO	0.00	0.69	0.81
deconstructSigs	0.41	0.75	0.86
MutationalPatterns	0.41	0.75	0.86
Sigfit	0.40	0.72	0.83
SigProfilerSS	0.43	0.74	0.83

Table 5. WGS data used to validate the *de novo* algorithm. The Hartwig Foundation and PCAWG datasets include cancer types such as breast cancer (BRCA), colon adenocarcinoma (COAD), kidney renal clear cell carcinoma (KIRC), pancreatic adenocarcinoma (PAAD), prostate adenocarcinoma (PRAD), and skin cutaneous melanoma (SKCM).

Source	Type	BRCA	KIRC	PAAD	PRAD	SKCM
Hartwig	WGS	792	612	186	402	327
PCAWG	WGS	214	145	237	286	107

Table 6. Comparison of signatures predicted by five *de novo* methods on skin tumour. Sparsity is measured as the fraction of cells in the matrix with a value = 0. Cosine similarity is measured as the mean of the reconstructed matrix and the original count matrix. Median per-patient correlation is measured as the median Pearson's correlation coefficient between the observed mutation section and the predicted mutational spectrum for each patient, indicating how well each method fits the observed mutations in individual patients. The prediction ratio is calculated between the total number of signatures, as detected with cosine similarity above 0.8, and the total number of signatures detected by each method. (The table header stands for; CS: Cosine Similarity; NR: Number signatures (K=$2 \cdots 18$), SigCos: Number of higher signature); Median: Median per-patient correlation

Method	CS	NR	MSE	SigCos	Pred Ratio	Sparsity	Median
StarSignDNA	0.99	4	74897	3	0.75	0,146	0.993
SigneR	0.99	18	6067	8	0,44	0.000	0.99
SigProfiler	0.99	5	26215	3	0.60	0.000	0.986
SUITOR	0.72	4	400693	3	0.75	0.000	0.719
SparseSignatures	0.99	11	82289	1	0.09	0,160	0.992

Table 7. Comparison of signatures predicted by five *de novo* methods on skin tumour. Sparsity is measured as the fraction of cells in the matrix with a value = 0. Cosine similarity is measured as the mean of the reconstructed matrix and the original count matrix. Median per-patient correlation is measured as the median Pearson's correlation coefficient between the observed mutation section and the predicted mutational spectrum for each patient, indicating how well each method fits the observed mutations in individual patients. The prediction ratio is calculated between the total number of signatures, as detected with cosine similarity above 0.8, and the total number of signatures detected by each method. (The table header stands for; CS: Cosine Similarity; NR: Number signatures (K=$2 \cdots 18$), SigCos: Number of higher signature); Median: Median per-patient correlation)

Method	CS	NR	MSE	SigCos	Ratio	Sparsity	Median
StarSignDNA	1.00	3	97837	3	1.00	0,041	0.994
SigneR	1.00	17	7713	8	0,47	0.000	0.99
SigProfiler	1.00	16	73239	8	0.50	0.018	0.98
SUITOR	0.88	3	606961	3	1.00	0.000	0.98
SparseSignatures	0.99	3	150344	2	0.67	0,131	0.995

Figure Legends

Figure 1. Performance comparison of six mutational signature refitting methods on simulated datasets. Each simulation included 100 samples with around 100 mutations each. The histograms represent the average relative contribution of each exposure, with vertical lines indicating the standard error. A green histogram alongside these shows the ground truth exposure values for comparison. The refitting process utilized SBS1, SBS2, SBS3, and SBS4 signatures from COSMIC as references. A) Only SBS1 has a true exposure of 1, while all other signatures have a true exposure of 0. B) Only SBS1 and SBS2 have a true exposure of 0.5 each, with all others at 0. C) SBS1, SBS2, and SBS3 each have a true exposure of 0.333, while SBS4 remains at 0. D) All four signatures have an equal true exposure of 0.25. For a more comprehensive analysis with different mutation counts (500, 1000, and 10,000 per sample), see Supplementary S1 Table.

Figure 2. Performance comparison of six mutational signature refitting methods on simulated datasets. Each simulation included 100 samples, with each

sample containing around 100 mutations (A), 500 mutations (B), or 1000 mutations (C). The histograms represent the average relative contribution of each exposure, with vertical lines indicating the standard error. A green histogram alongside these are the ground truth exposure values for comparison. The refitting process utilized SBS7d, SBS10b, SBS13, SBS17b, and SBS28 signatures from COSMIC as references. (D) shows the average fitting error (SimE) across 100 cohorts from each simulation with 100, 500, and 1000 mutations (SimE(100), SimE(500), and SimE(1000))

Figure 3. Comparison of F1 scores of six refitting methods using SimE(100), SimE(500) and SimE(1000) datasets. The F1 scores were used to evaluate the precision and sensitivity of the signature assignment across the simulations with the threshold exposure of 0.08 (i.e. elements below 0.08 are considered to be zero).

Figure 4. Performance comparison of six mutational signature refitting methods on simulated datasets. Each simulation included 100 samples, with each sample containing around 100 mutations (A), 500 mutations (B), or 1000 mutations (C). The histograms represent the average relative contribution of each exposure, with error bars indicating the standard error. A green histogram alongside these are the ground truth exposure values for comparison. The refitting process utilized SBS3, SBS5, SBS40c, SBS92, and SBS94 signatures from COSMIC as references. (D) shows the average fitting error (SimF) across 100 cohorts from each simulation with 100, 500, and 1000 mutations (SimF(100), SimF(500), and SimF(1000)).

Figure 5. Comparison of F1 scores of six refitting methods using SimF(100), SimF(500) and SimF(1000) datasets. The F1 scores were used to evaluate the precision and sensitivity of the signature assignment across the simulations with the threshold exposure of 0.08 (i.e. elements below 0.08 are considered to be zero).

Figure 6. The F1 score was used to evaluate the accuracy of signature assignment across six refitting methods. The number of samples for each dataset is 100 with different numbers of mutations. An F1 score threshold of 0.08 was used to evaluate the accuracy. This means elements below 0.08 were considered zero for the signature assignment

Figure 7. This figure shows the performance of StarSignDNA on 33 TCGA cancer types listed alphabetically along the top x-axis. The size of each dot represents the proportion of signatures within a specific cancer type that exhibits a mutational signature identified by StarSignDNA. The bottom x-axis indicates the number of samples analyzed for each cancer type. A separate panel legend identifies the "active signatures" (those contributing more than 6% of the mutations) in each cancer type using different methods. All 86 COSMIC v3.4 signatures were used as references for the analysis.

Figure 8. Comparison of mutational refitting algorithm for a single sample. The y-axis represents the relative contribution of the exposure, the bar extends from the 2.5% percentile to the 97.5% percentile of the bootstrapped exposure levels obtained from 200 bootstraps, and the x-axis represents the signature exposures.

Figure 9. Comparison of StarSignDNA with Other Methods on Simulated Data. A) Number of Signatures Detected (K): This bar chart shows how often each method selected a specific number of signatures (K) across the 50 simulations. The

x-axis represents the possible values of K, and the y-axis shows the frequency at which each method chose a particular K. All methods were run on 50 simulated datasets with a ground truth value of K=4 (i.e., the actual number of underlying processes in the data). B) Residual Error: This box plot compares the residual error achieved by each method across the 50 simulations. The residual error is measured using the Mean Squared Error (MSE) between the reconstructed and original count matrices. Lower MSE indicates better reconstruction. C) Cosine Similarity: This box plot shows the cosine similarity between the reconstructed signatures and the original input signatures for each method across the 50 simulations. A higher cosine similarity value signifies a more accurate reconstruction of the original signatures. D) Sparsity of Signatures: This box plot compares the sparsity of the signatures generated by each method across the simulations. Sparsity is measured as the proportion of elements in the signature matrix with values below a threshold of 10^3. Higher sparsity indicates a higher number of zeros within the signatures. The data underlying these analyses can be found in (S6 Table).

Figure 10. *In silico* evaluation of StarSignDNA and other methods. The number of replicates in which a given signature (of a total of 4 signatures) will be detected by each method.

Figure 11. Mutational signature results of five cancer types in PCAWG. The cosine similarities are between *de novo* signatures and COSMIC signatures. Only the matched pairs are shown (with cosine similarity ¿ 0.8). The higher the cosine similarity, the better the match to the COSMIC signature profile. The cosine similarity equivalent to one denotes a perfect match. The value of K represents the number of *de novo* signatures discovered by each method.

Figure 12. Mutational signature results of five metastatic cancer types from Hartwig data. The cosine similarities between *de novo* signatures and COSMIC signatures are shown. Only the matched pairs are shown (with cosine similarity ¿ 0.8). The higher the cosine similarity, the better the match to a COSMIC signature profile. A cosine similarity equivalent to one denotes a perfect match. The value of K represents the number of *de novo* signatures discovered by each method.

Figure 13. Comparison of *de novo* Signature Quality by Algorithm. A) PCAWG Skin Cancer compares the quality of *de novo* signatures identified by different algorithms for PCAWG skin cancer data. The analysis considers a set of 11 mutational signatures: SBS1, SBS2, SBS5, SBS7a, SBS7b,SBS7c, SBS7d, SBS13, SBS17a, SBS17b, and SBS40. B) PCAWG Breast Cancer compares the quality of *de novo* signatures identified by different algorithms for PCAWG breast cancer data. The analysis considers a set of 12 mutational signatures: SBS1, SBS2, SBS3, SBS5, SBS8, SBS13, SBS17a, SBS17b,SBS18, SBS37, SBS40, and SBS41.

Supporting information

Supporting Figure 1. COSMIC version 3.4 reference signature Cosine Similarity distribution . The red points highlight the distribution and correlation of mutation frequencies with an average cosine similarity of 0.32 and a maximum of 0.6 for the flat signatures: SBS3, SBS5, SBS40c, SBS92, and SBS94. The blue points highlight the distribution and correlation of mutation frequencies with an average cosine similarity of 0.29 and a maximum of 0.6 for the common signatures: SBS7d, SBS10b, SBS13, SBS17b, and SBS28. The black points represent the rest of the signatures. This

plot helps visualize the relationship between mean (xmean) and maximum (xmax) mutation frequencies for various SBS signatures.

Supporting Figure 2. Refitting algorithm running time. Running time of StarSignDNA, deconstructSigs, MutationalPatterns, sigLASSO and sigfit at different numbers of signatures simulated with 100 sample of 100 mutation per sample.

Supporting Figure 3. Hyperparameter λ obtained using cross-validation for WES and WGS.

Supporting Figure 4. This figure shows the performance of sigLASSO on 33 TCGA cancer types listed alphabetically along the top x-axis. The size of each dot represents the proportion of signatures within a specific cancer type that exhibits a mutational signature identified by sigLASSO. The bottom x-axis indicates the number of samples analyzed for each cancer type. A separate panel legend identifies the *active signatures* (those contributing more than 6% of the mutations) in each cancer type using different methods. All 86 COSMIC v3.4 signatures were used as references for the analysis.

Supporting Figure 5. This figure shows the performance of deconstructSigs on 33 TCGA cancer types listed alphabetically along the top x-axis. The size of each dot represents the proportion of signatures within a specific cancer type that exhibits a mutational signature identified by deconstructSigs. The bottom x-axis indicates the number of samples analyzed for each cancer type. A separate panel legend identifies the *active signatures* (those contributing more than 6% of the mutations) in each cancer type using different methods. All 86 COSMIC v3.4 signatures were used as references for the analysis.

S1 Table. Exposure matrices of six mutational signature refitting methods on simulated datasets. The exposure matrix of the second category of the simulated dataset uses common signatures SBS7d, SBS10b, SBS13, SBS17b, and SBS28 as a reference signature with 100 samples, each having 100 mutations per sample.

S2 Table. Exposure matrices of six mutational signature refitting methods on simulated datasets. The exposure matrix of the second category of the simulated dataset uses common signatures SBS7d, SBS10b, SBS13, SBS17b, and SBS28 as a reference signature with 100 samples, each having 500 mutations per sample.

S3 Table. Exposure matrices of six mutational signature refitting methods on simulated datasets. The exposure matrix of the second category of the simulated dataset uses common signatures SBS7d, SBS10b, SBS13, SBS17b, and SBS28 as a reference signature with 100 samples, each having 1000 mutations per sample.

S4 Table. Exposure matrices of six mutational signature refitting methods on simulated datasets. The exposure matrix of the second category of the simulated dataset with 100 samples, each having 100 mutations per sample. All 86 COSMIC v3.4 signatures were used as references for the analysis.

S5 Table. Exposure matrices of six mutational signature refitting methods on simulated datasets. The exposure matrix of the second category of the simulated dataset with 100 samples, each having 500 mutations per sample. All 86 COSMIC v3.4 signatures were used as references for the analysis.

S6 Table. Exposure matrices of six mutational signature refitting methods on simulated datasets. The exposure matrix of the second category of the simulated dataset with 100 samples, each having 1000 mutations per sample. All 86 COSMIC v3.4 signatures were used as references for the analysis.

S7 Table. Exposure matrices of six mutational signature refitting methods on simulated datasets. The exposure matrix of the third category of the simulated dataset uses common signatures SBS3, SBS5, SBS40c, SBS92, and SBS94 as a reference signature with 100 samples, each having 100 mutations per sample.

S8 Table. Exposure matrices of six mutational signature refitting methods on simulated datasets. The exposure matrix of the third category of the simulated dataset uses common signatures SBS3, SBS5, SBS40c, SBS92, and SBS94 as a reference signature with 500 samples, each having 100 mutations per sample.

S9 Table. Exposure matrices of six mutational signature refitting methods on simulated datasets. The exposure matrix of the third category of the simulated dataset uses common signatures SBS3, SBS5, SBS40c, SBS92, and SBS94 as a reference signature with 1000 samples, each having 100 mutations per sample.

S10 Table. Exposure matrices of six mutational signature refitting methods on simulated datasets. The exposure matrix of the third category of the simulated dataset with 100 samples, each having 100 mutations per sample. All 86 COSMIC v3.4 signatures were used as references for the analysis.

S11 Table. Exposure matrices of six mutational signature refitting methods on simulated datasets. The exposure matrix of the third category of the simulated dataset with 100 samples, each having 500 mutations per sample. All 86 COSMIC v3.4 signatures were used as references for the analysis.

S12 Table. Exposure matrices of six mutational signature refitting methods on simulated datasets. The exposure matrix of the third category of the simulated dataset with 100 samples, each having 1000 mutations per sample. All 86 COSMIC v3.4 signatures were used as references for the analysis.

S13 Table. Exposure matrices of six mutational signature refitting methods on simulated datasets. The exposure matrix of the third category of the simulated downsample dataset with 100 samples, each having 100 mutations per sample. All 86 COSMIC v3.4 signatures were used as references for the analysis.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 801133.

Author Contributions

Conceptualization: Christian Domilongo Bope, Sigve Nakken, Ole Christian Lingjærde, and Eivind Hovig.

Developed the mathematical framework: Christian Domilongo Bope and Ole Christian Lingjærde.

Software: Christian Domilongo Bope, Knut D. Rand and Sumana Kalyanasundaram.

Performed the simulations, benchmarking, and results validation: Christian Domilongo Bope and Sumana Kalyanasundaram.

Interpretation of results and manuscript preparation: Christian Domilongo Bope, Sumana Kalyanasundaram, Sigve Nakken, Ole Christian Lingjærde and Eivind Hovig.

All the authors critically reviewed and approved the final manuscript.

References

1. Petljak M and Alexandrov LB. Understanding mutagenesis through delineation of mutational signatures in human cancer. *Carcinogenesis*, 37(6), 2016.

2. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, and et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463), 2013.

3. Brady SW, Gout AM, and Zhang J. Therapeutic and prognostic insights from the analysis of cancer mutational signatures. *Trends in Genetics*, 38(2):194–208, 2022.

4. Steele CD, Abbasi A, Islam SMA, Bowes AL, Khandekar A, and et al. Signatures of copy number alterations in human cancer. *Nature*, 606(7916):984–991, 2022.

5. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, and et al. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, 2018.

6. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, and et al. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep*, 3(1):246–59, 2013.

7. Baez-Ortega A and Gori K. Computational approaches for discovery of mutational signatures in cancer. *Briefing Bioinformatics*, 20(1):77–88, 2019.

8. Rosales RA, Drummond RD, Valieris R, Dias-Neto E, and da Silva IT. signer: an empirical bayesian approach to mutational signature discovery. *Bioinformatics*, 33:8–16, 2017.

9. Kim J, Mouw KW, Polak P, Braunstein LZ, Kamburov A, and et al. Somatic ercc2 mutations are associated with a distinct genomic signature in urothelial tumors. *National Genetics*, 48(6):600–606, 2017.

10. Fantini D, Huang S, Asara JM, Bagchi S, and Raychaudhuri P. Chromatin association of xrcc5/6 in the absence of dna damage depends on the xpe gene product ddb2. *Molecular Biology of the Cell*, 28(6):192–200, 2017.

11. Gehring JS, Fischer B, Lawrence M, and Huber W. Somaticsignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics*, 31(22):3673–5, 2015.

12. Lal A, Liu K, Tibshirani R, Sidow A, and Ramazzotti D. De novo mutational signature discovery in tumor genomes using sparsesignatures. *PLoS Comput Biol*, 17(6), 2021.

13. Helleday T, Eshtad S, and Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet*, 15(9):585–98, 2014.

14. Fischer A, Illingworth CJ, Campbell PJ, and Mustonen V. Emu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biology*, 14, 2013.

15. Shiraishi Y, Tremmel G, Miyano S, and Stephens M. A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLoS Genet*, 11(12), 2015.

16. Rosenthal R, McGranahan N, Herrero J, Barry ST, and Swanton C. deconstructsigs: delineating mutational processes in single tumors distinguishes dna repair deficiencies and patterns of carcinoma evolution. *Genome Biology*, 17:31, 2017.

17. Otlu B and Alexandrov LB. Evaluating topography of mutational signatures with sigprofilertopography. *bioRxiv Preprint*, 2024.

18. Li S, Crawford FW, and Gerstein MB. Using siglasso to optimize cancer mutation signatures jointly with sampling likelihood. *Nature Communications*, 11(1), 2020.

19. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, and et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, 2016.

20. Lee D, Wang D, Yang XR, Shi J, Landi MT, and Zhu B. Suitor: Selecting the number of mutational signatures through cross-validation. *PLoS Comput Biol*, 18(4), 2022.

21. Kevin Gori and Adrian Baez-Ortega. sigfit: flexible bayesian inference of mutational signatures. *bioRxiv*, page 372896, 2018.

22. Kim H and Park H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2021.

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

24. Rustad EH, Nadeu F, Angelopoulos N, Ziccheddu B, Bolli N, and et al. mmsig: A fitting approach to accurately identify somatic mutational signatures in hematological malignancies. *Communications Biology*, 4(424):1–12, 2021.

25. Omichessan H, Severi G, and 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), 2019.

26. Jelle J. Goeman. L1 penalized estimation in the cox proportional hazards model. *Biomedical Journal*, 52(2):70–84, 2010.

27. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

28. Tibshirani R. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):386–395, 1997.

29. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on Artificial intelligence*, volume 2, pages 386–395, 1995.

30. Breiman L and Spector P. Submodel selection and evaluation in regression: the x-random case. *International Statistical Review / Revue Internationale de Statistique*, 60(3):291–319, 1992.

Figure 1

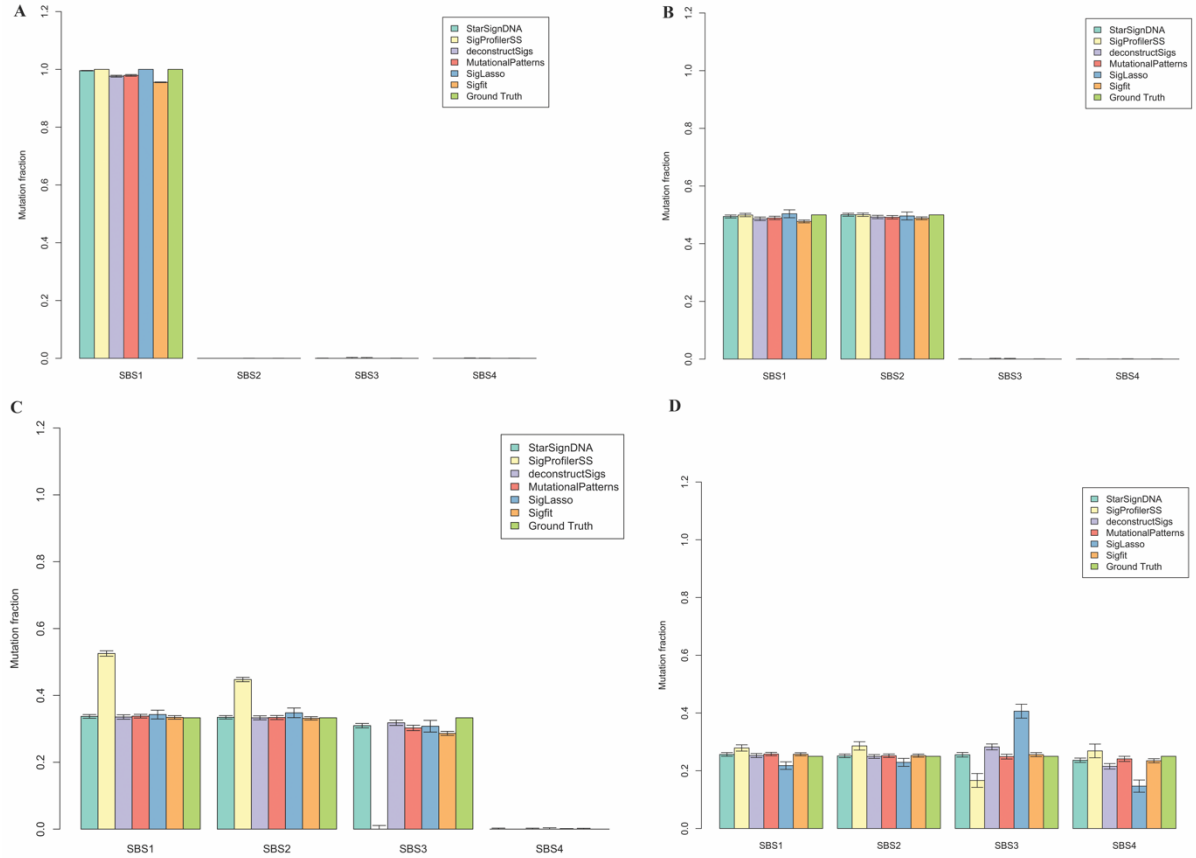

A
B
C
D
Mutation fraction
0.0
0.2
0.4
0.6
0.8
1.0
1.2
SBS1
SBS2
SBS3
SBS4
StarSignDNA
SigProfilerSS
deconstructSigs
MutationalPatterns
SigLasso
Sigfit
Ground Truth

Figure 2

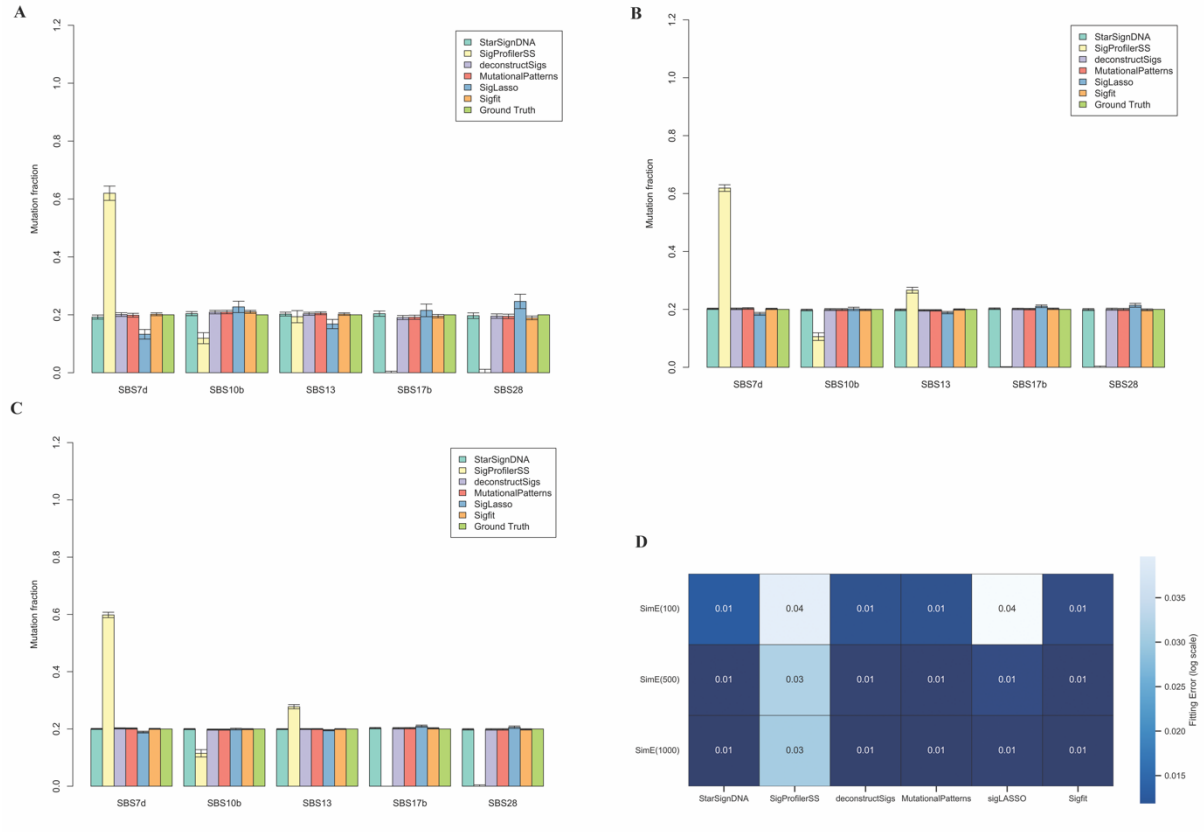


Figure 3

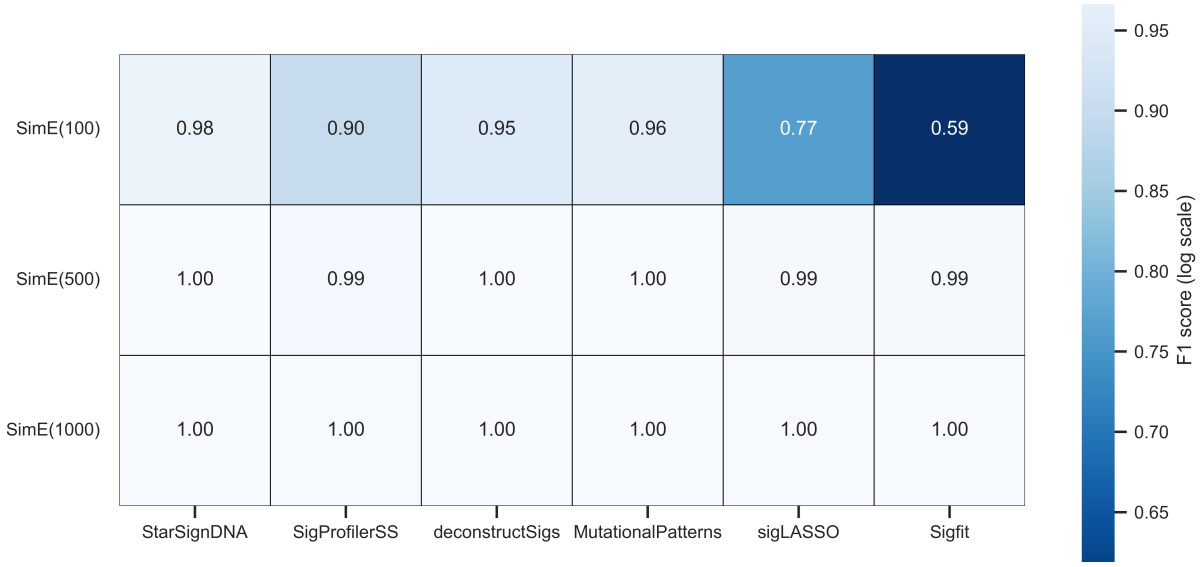


Figure 4

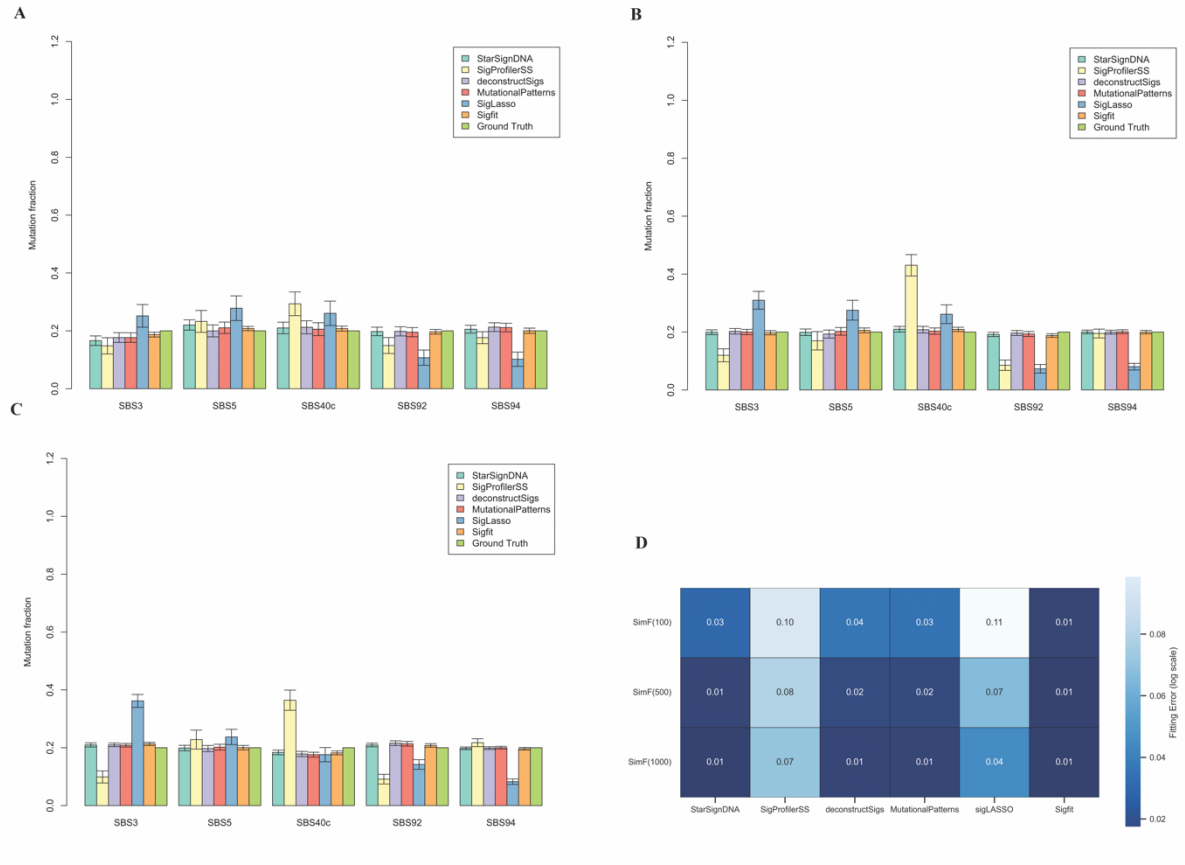

A
B
C
D
StarSignDNA
SigProfilerSS
deconstructSigs
MutationalPatterns
SigLasso
Sigfit
Ground Truth
Mutation fraction
SBS3
SBS5
SBS40c
SBS92
SBS94
SimF(100)
SimF(500)
SimF(1000)
0.03
0.10
0.04
0.03
0.11
0.01
0.01
0.08
0.02
0.02
0.07
0.01
0.01
0.07
0.01
0.01
0.04
0.01
sigLASSO
Fitting Error (log scale)

Figure 5

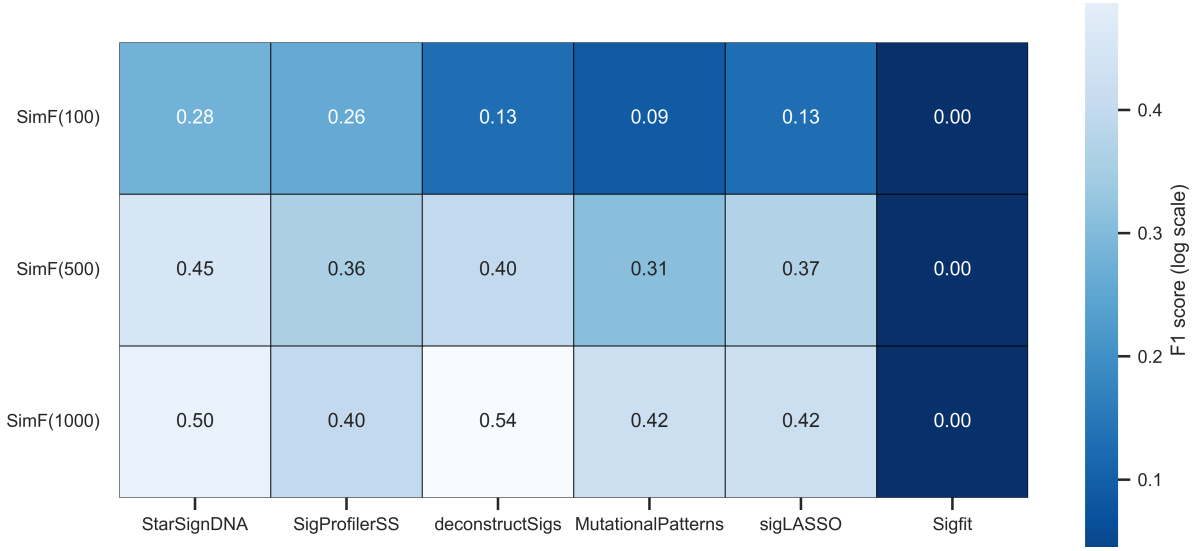


Figure 6

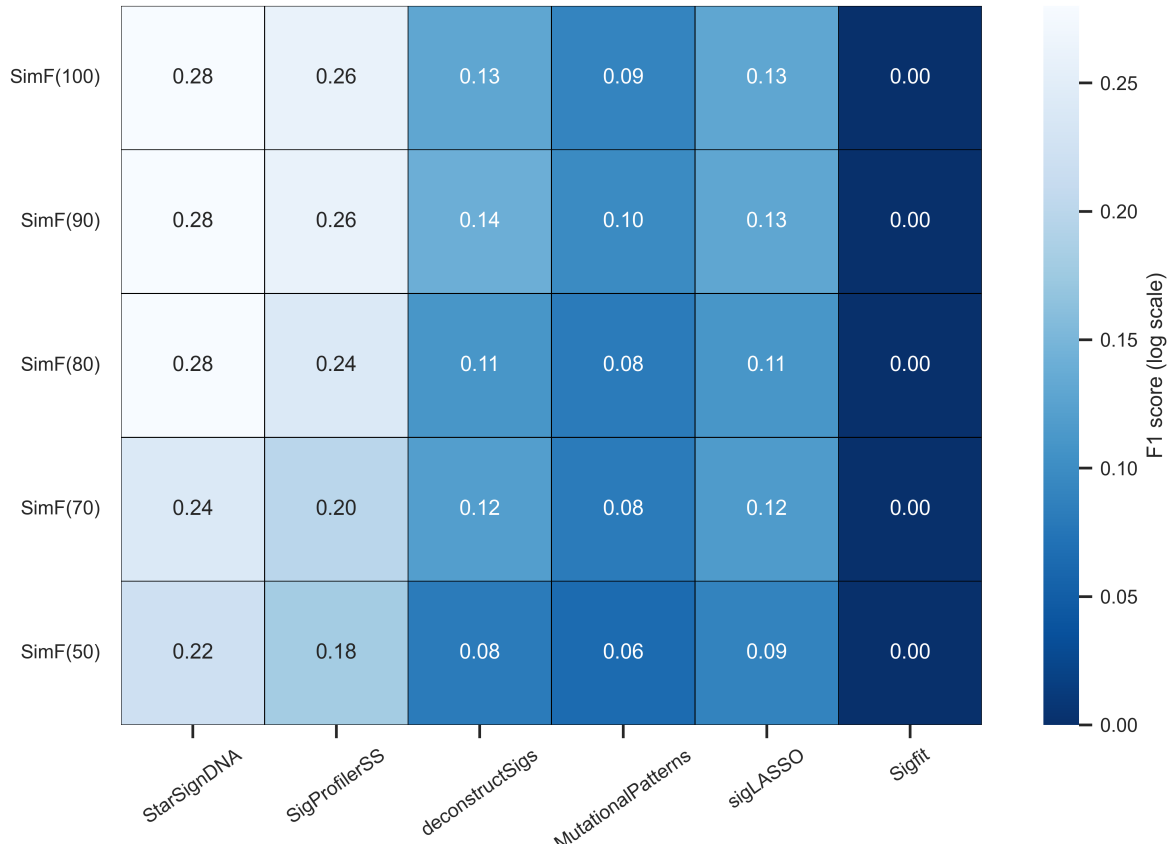


Figure 7

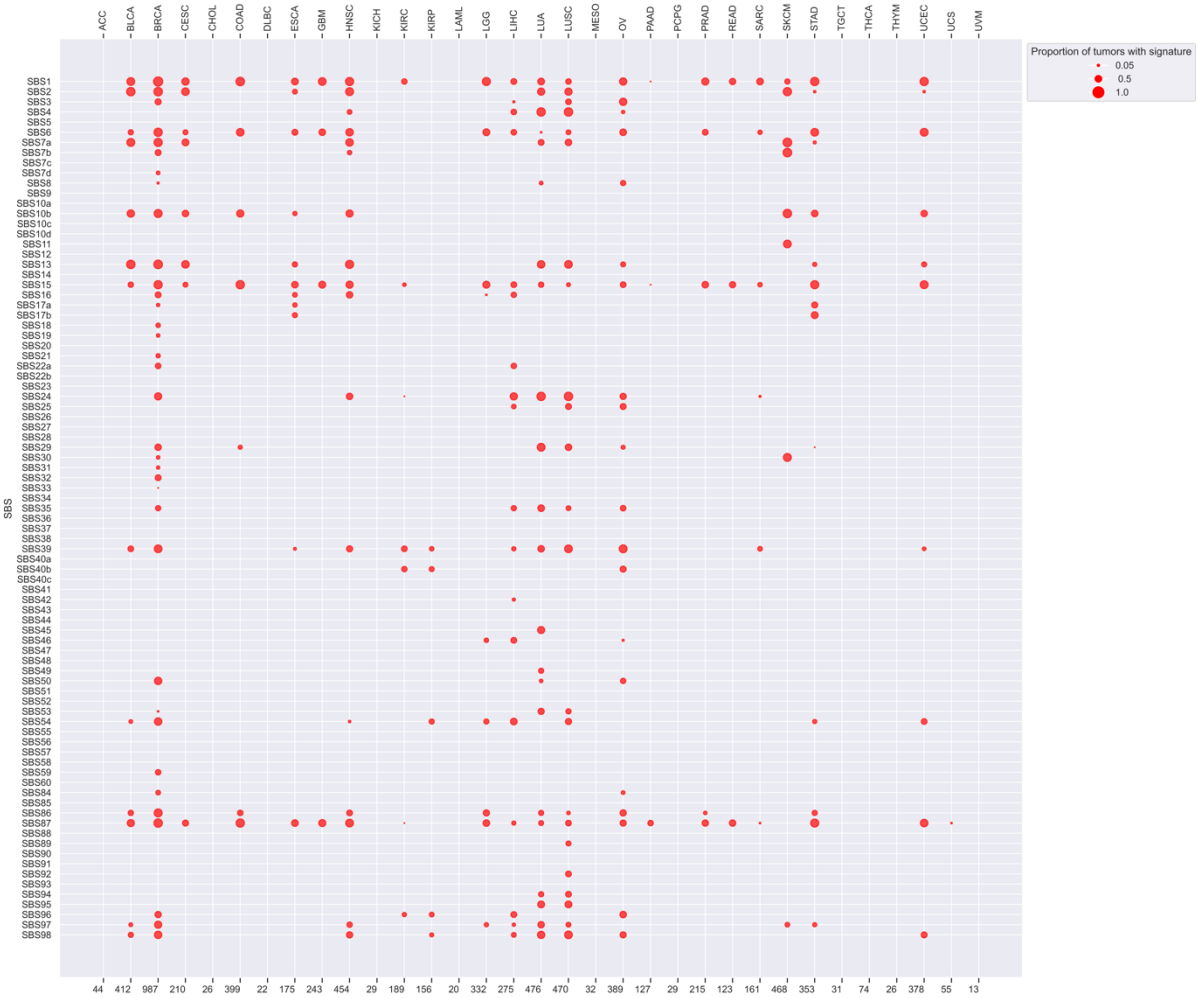


Figure 8

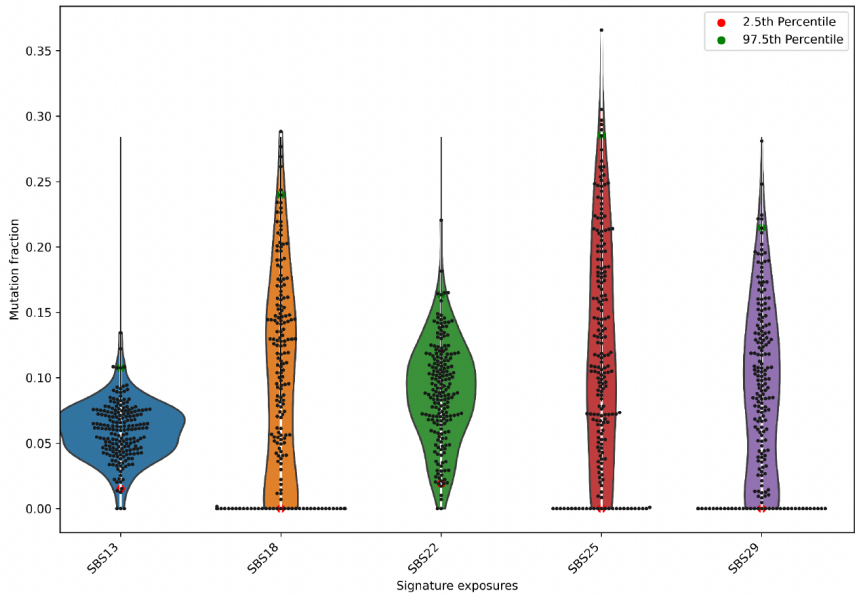


Figure 9

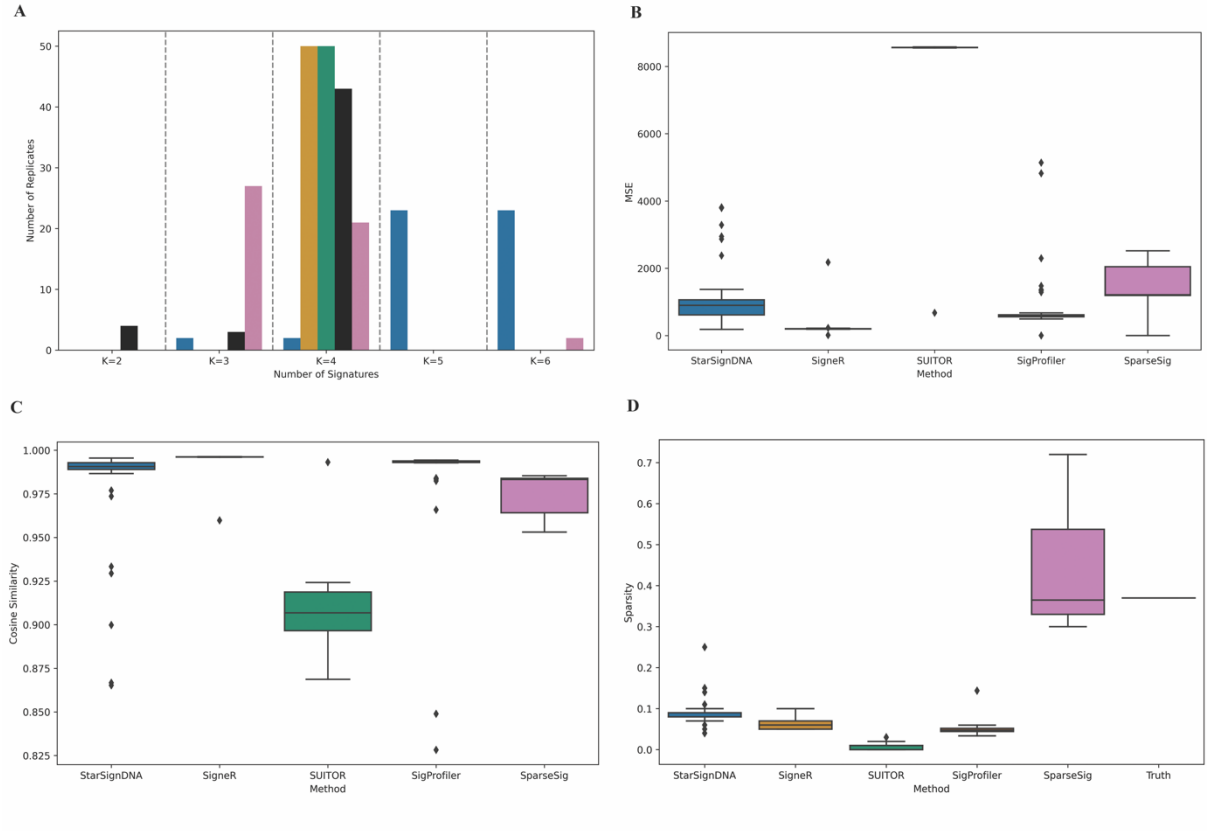


Figure 10

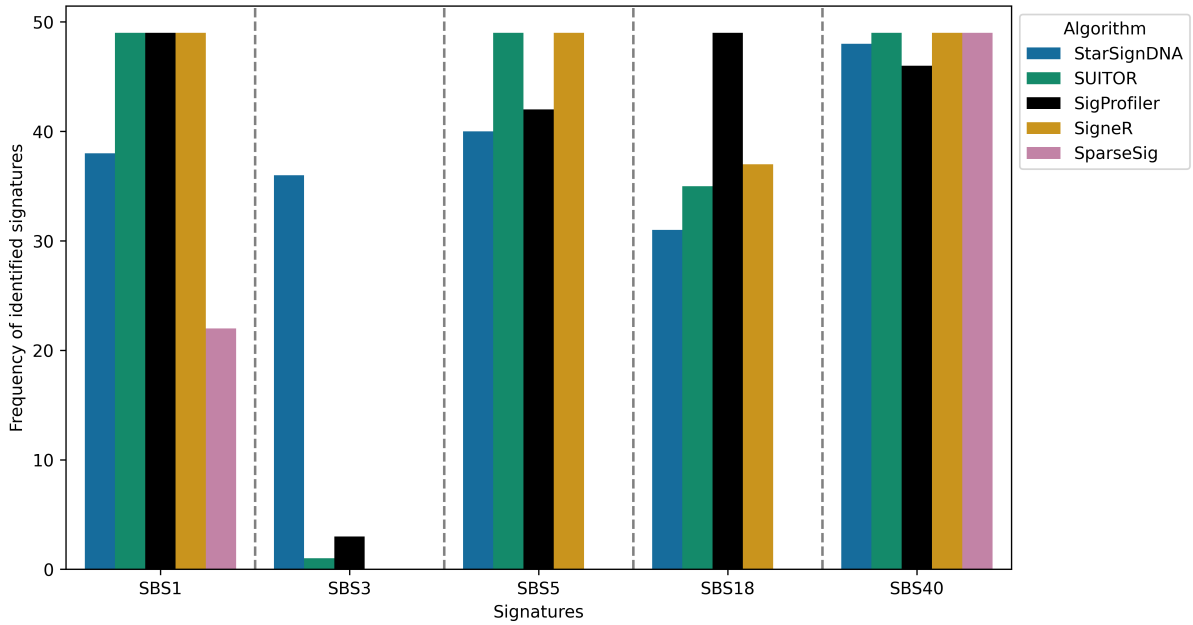


Figure 11

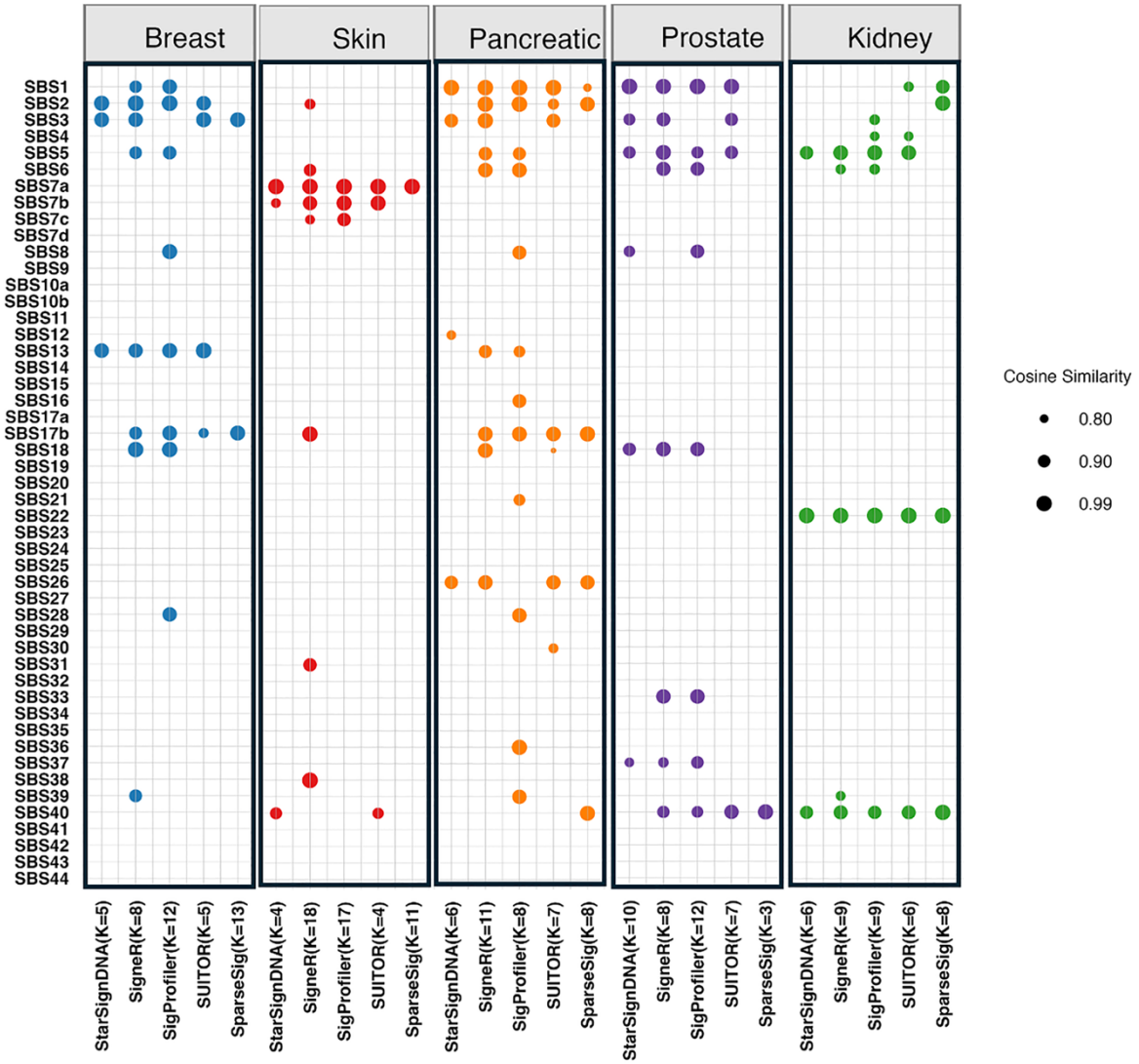

Breast
Skin
Pancreatic
Prostate
Kidney
SBS1
SBS2
SBS3
SBS4
SBS5
SBS6
SBS7a
SBS7b
SBS7c
SBS7d
SBS8
SBS9
SBS10a
SBS10b
SBS11
SBS12
SBS13
SBS14
SBS15
SBS16
SBS17a
SBS17b
SBS18
SBS19
SBS20
SBS21
SBS22
SBS23
SBS24
SBS25
SBS26
SBS27
SBS28
SBS29
SBS30
SBS31
SBS32
SBS33
SBS34
SBS35
SBS36
SBS37
SBS38
SBS39
SBS40
SBS41
SBS42
SBS43
SBS44
StarSignDNA(K=5)
SigneR(K=8)
SigProfiler(K=12)
SUITOR(K=5)
SparseSig(K=13)
StarSignDNA(K=4)
SigneR(K=18)
SigProfiler(K=17)
SUITOR(K=4)
SparseSig(K=11)
StarSignDNA(K=6)
SigneR(K=11)
SigProfiler(K=8)
SUITOR(K=7)
SparseSig(K=8)
StarSignDNA(K=10)
SigneR(K=8)
SigProfiler(K=12)
SUITOR(K=7)
SparseSig(K=3)
StarSignDNA(K=6)
SigneR(K=9)
SigProfiler(K=9)
SUITOR(K=6)
SparseSig(K=8)
Cosine Similarity
0.80
0.90
0.99

Figure 12

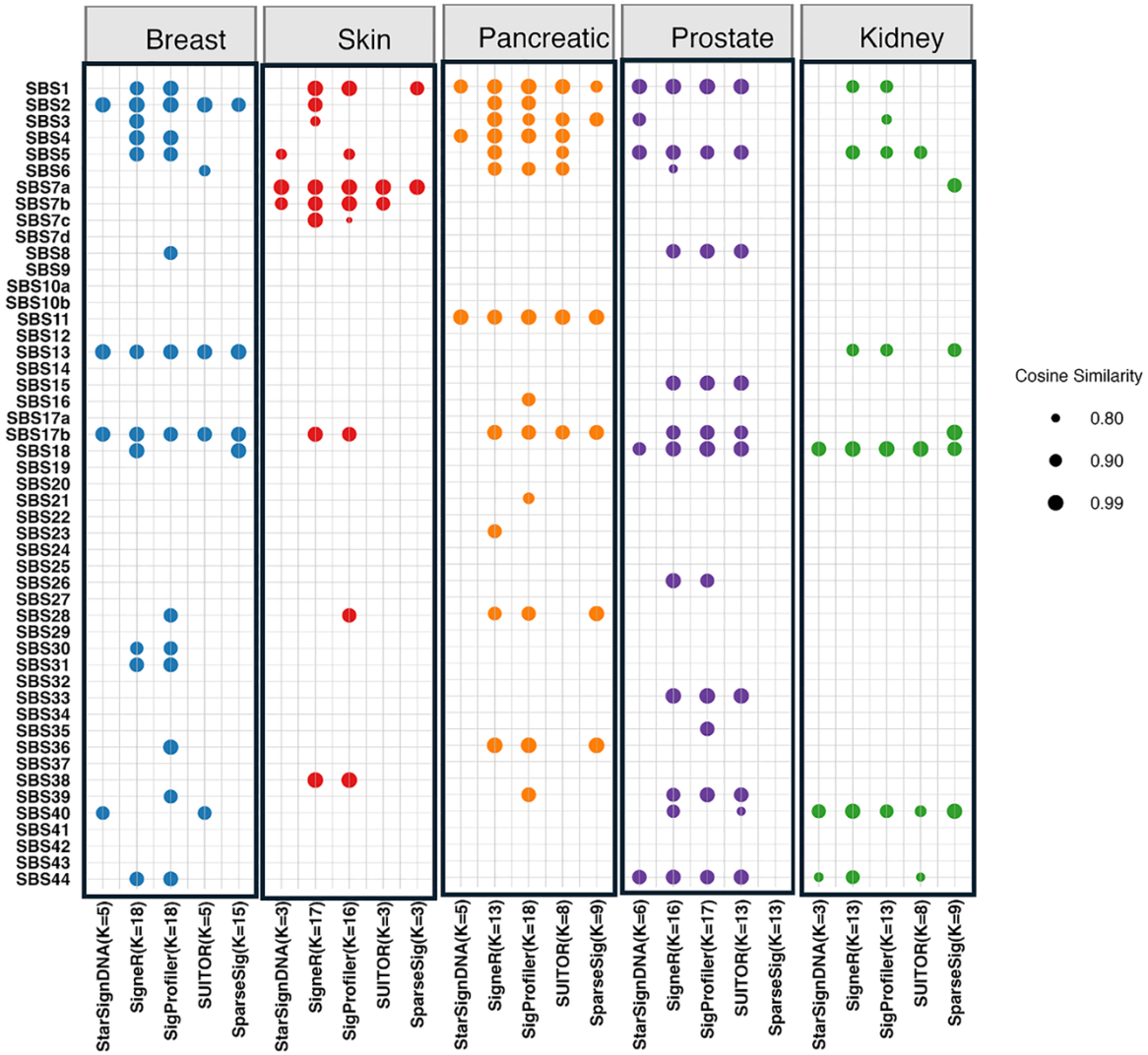
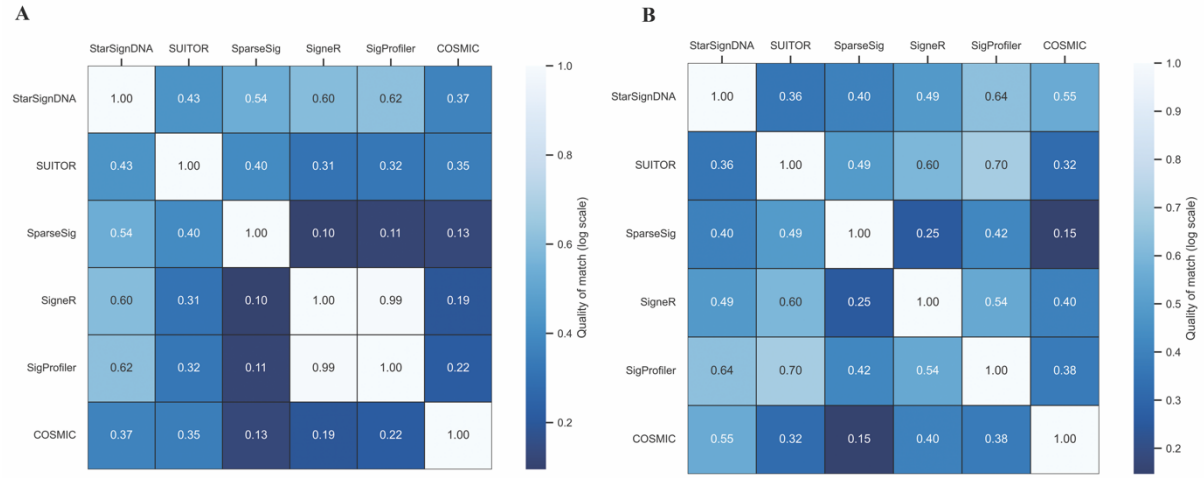
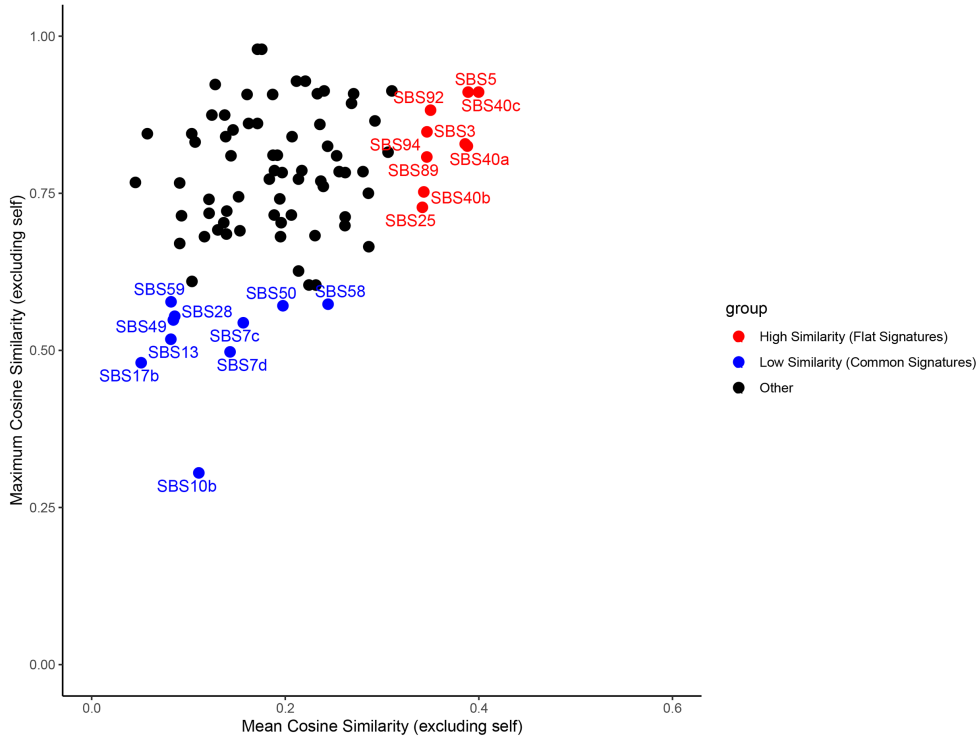

Breast
Skin
Pancreatic
Prostate
Kidney
SBS1
SBS2
SBS3
SBS4
SBS5
SBS6
SBS7a
SBS7b
SBS7c
SBS7d
SBS8
SBS9
SBS10a
SBS10b
SBS11
SBS12
SBS13
SBS14
SBS15
SBS16
SBS17a
SBS17b
SBS18
SBS19
SBS20
SBS21
SBS22
SBS23
SBS24
SBS25
SBS26
SBS27
SBS28
SBS29
SBS30
SBS31
SBS32
SBS33
SBS34
SBS35
SBS36
SBS37
SBS38
SBS39
SBS40
SBS41
SBS42
SBS43
SBS44
StarSignDNA(K=5)
SigneR(K=18)
SigProfiler(K=18)
SUITOR(K=5)
SparseSig(K=15)
StarSignDNA(K=3)
SigneR(K=17)
SigProfiler(K=16)
SUITOR(K=3)
SparseSig(K=3)
StarSignDNA(K=5)
SigneR(K=13)
SigProfiler(K=18)
SUITOR(K=8)
SparseSig(K=9)
StarSignDNA(K=6)
SigneR(K=16)
SigProfiler(K=17)
SUITOR(K=13)
SparseSig(K=13)
StarSignDNA(K=3)
SigneR(K=13)
SigProfiler(K=13)
SUITOR(K=8)
SparseSig(K=9)
Cosine Similarity
0.80
0.90
0.99

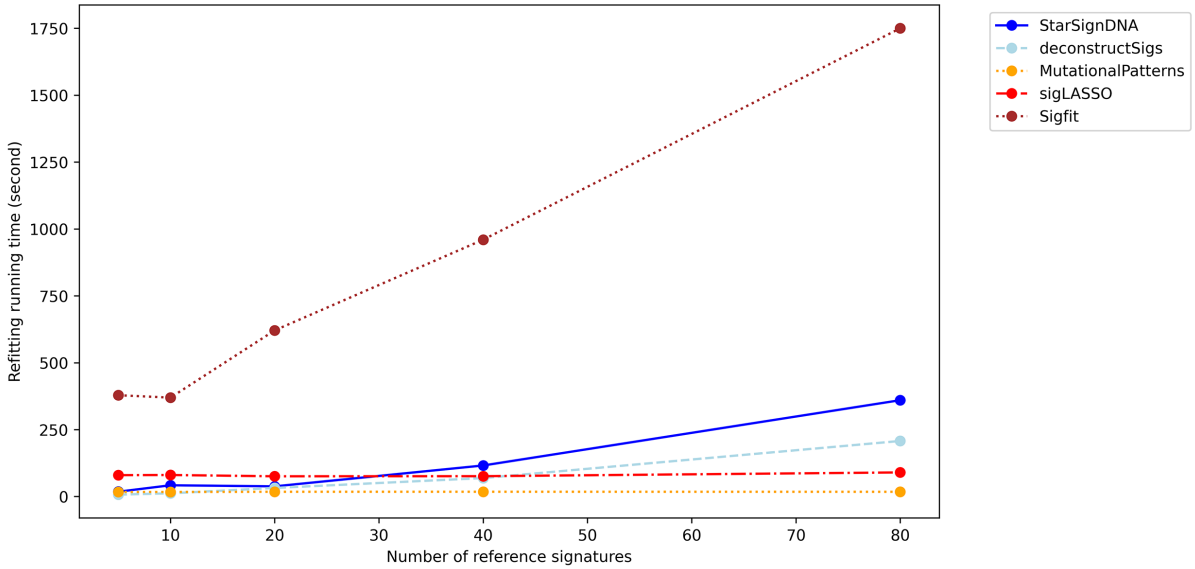
Figure 13



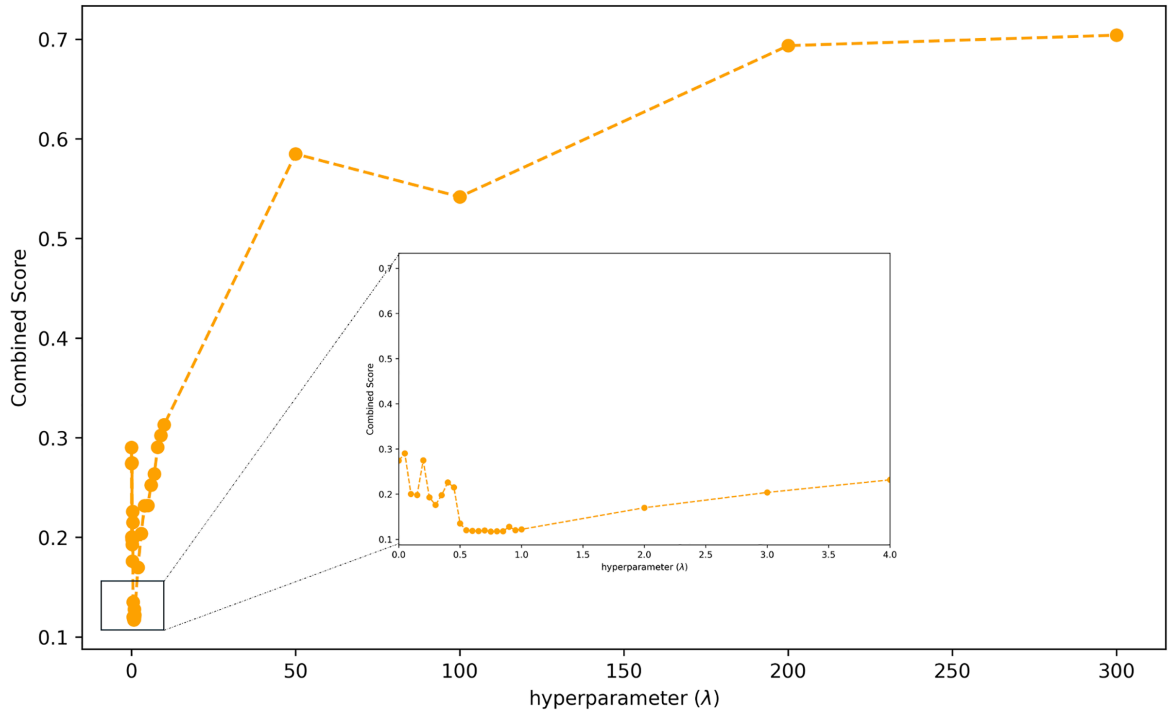
Supporting Figure 1



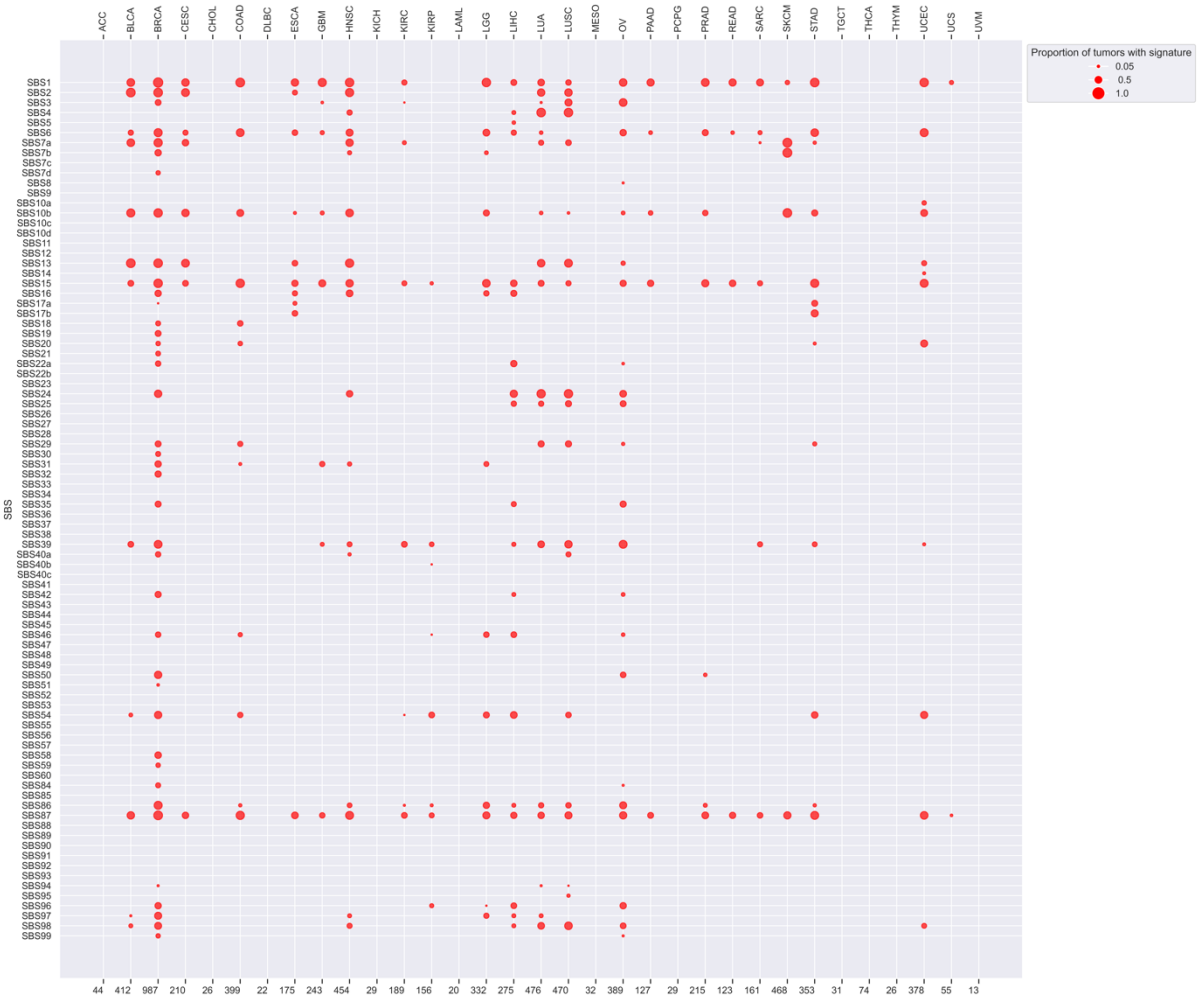
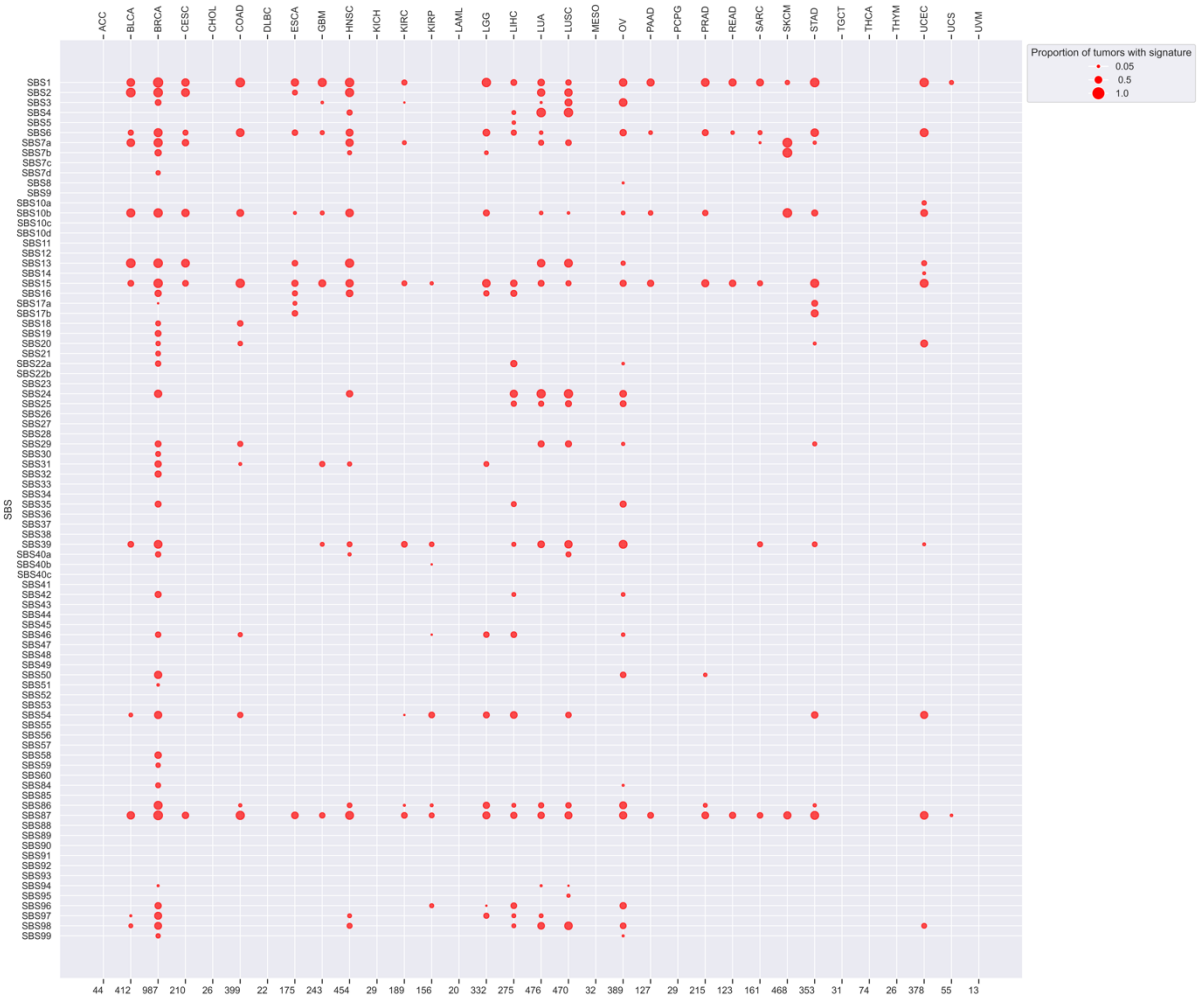
Supporting Figure 2



Supporting Figure 3



Supporting Figure 4



Supporting Figure 5