De Novo Mutational Signature Discovery in Tumor Genomes using SparseSignatures

Avantika Lal[1,*], Keli Liu[2], Robert Tibshirani[2,3], Arend Sidow[1,4] and Daniele Ramazzotti[1,5,#]

[1]Department of Pathology, Stanford University; [2]Department of Statistics, Stanford University; [3]Department of Biomedical Data Science, Stanford University; [4]Department of Genetics, Stanford University; [5]Department of Computer Science, Stanford University.

[*]Present address: NVIDIA Corporation, 2788 San Tomas Expy, Santa Clara, CA 95051, USA.

[#]Present address: Department of Medicine and Surgery, University of Milan-Bicocca, Monza, Italy.

Corresponding author: Daniele Ramazzotti, Department of Medicine and Surgery, University of Milan-Bicocca, Monza, Italy. daniele.ramazzotti@unimib.it

Running title: Mutational Signature Discovery by SparseSignatures

Keywords: Mutational Signatures, Cancer Genomics, Mutagenesis Mechanisms

Abstract

Cancer is the result of mutagenic processes that can be inferred from tumor genomes by analyzing rate spectra of point mutations, or "mutational signatures". Here we present SparseSignatures, a novel framework to extract signatures from somatic point mutation data. Our approach incorporates DNA replication error as a background, employs regularization to reduce noise in non-background signatures, uses cross-validation to identify the number of signatures, and is scalable to large datasets. We show that SparseSignatures outperforms current state-of-the-art methods on simulated data using standard metrics. We then apply SparseSignatures to whole genome sequences of 147 tumors from pancreatic cancer, discovering 8 signatures in addition to the background.

Introduction

Cancer is caused by somatic mutations in genes that control cellular growth and division (Vogelstein et al. 2013). The chance of developing cancer is massively elevated if mutagenic processes (e.g., defective DNA repair, environmental mutagens) increase the rate of somatic mutations. Due to the specificity of molecular lesions caused by such processes, and the specific repair mechanisms deployed by the cell to mitigate the damage, mutagenic processes generate characteristic point mutation rate spectra ('signatures') (Alexandrov et al. 2013). These signatures can indicate which mutagenic processes are active in a tumor, reveal biological differences between cancer subtypes, and may be useful markers for therapeutic response (Wang et al. 2018a).

Signatures are discovered by identifying common patterns across tumors based on counts of mutations and their sequence context. The original signature discovery method was based on Non-Negative Matrix Factorization (NMF) (Alexandrov et al. 2013). While other approaches have been considered (Gehring et al. 2015; Shiraishi et al. 2015; Fischer et al. 2013), NMF-based methods are by far

the most widely used (Bolli et al. 2014; Schulze et al. 2015; Nik-Zainal et al. 2016) and have resulted in a commonly used catalog of 30 signatures across human cancers (Alexandrov et al. 2015), available in the COSMIC version 2 database (https://cancer.sanger.ac.uk/cosmic/signatures_v2). A recent study (Alexandrov et al. 2020) using two NMF-based methods presented higher numbers (49 and 60) of putative signatures, which has now been incorporated into version 3 of the COSMIC database (https://cancer.sanger.ac.uk/cosmic/signatures).

While some reported signatures have been associated with mutagenic processes (Alexandrov et al. 2016; Nik-Zainal et al. 2015; Helleday et al. 2014), careful examination reveals that several reported signatures are highly similar, suggesting overfitting rather than distinct mutagenic processes. In addition, there are several 'flat' signatures of uncertain origin (non-specific signatures that include mutations of all types and sequence contexts), and many signatures appear to be distorted by low levels of background noise. As an example, one may consider SBS40 in COSMIC version 3, whose etiology is unclear and which has many features in common with SBS5 (Alexandrov et al. 2020). Another example is represented by the four similar signatures in COSMIC version 2 that are attributed to defective DNA mismatch repair (signatures 6, 15, 20, and 26), which share common features and are not clearly separated. Such uncertainty complicates the task of understanding which signatures are active in different patients. These observations are consistent with critical weaknesses in current signature discovery studies:

(1) State-of-the-art NMF-based methods aim to minimize the residual error after fitting the dataset with the discovered signatures (Alexandrov et al. 2013, Gehring et al. 2015), in an effort to fit the dataset perfectly. Consequently they may overfit by including stochastic noise in the dataset as part of the signatures, or by producing multiple similar signatures for the same underlying process. This problem is exacerbated by the relatively low number of samples (hundreds or thousands) available to most mutational signature discovery studies. LASSO regularization has been shown to improve estimation in high dimensional problems when the sample size is small

relative to the number of parameters (Tibshirani 1996). A method that applies LASSO regularization on the signatures would help alleviate the aforementioned drawbacks by favoring well-differentiated signatures with low background noise, in addition to minimizing residual error.

Variants of NMF that incorporate regularization are available and have been used in other domains (Pascual-Montano et al. 2006, Kim et al. 2007, the NNLM R package on CRAN at https://cran.r-project.org/web/packages/NNLM/index.html), and a few recent studies (Covington et al. 2016; Goncearenco et al. 2017) have attempted to apply these methods to signature discovery.

(2) No method incorporates the natural background of 'standard' replication error and repair processes, which occur in the normal course of cell division both in the germline and in somatic cells, including those of a tumor (Ledford 2017). Since we expect it to be present in all samples, and since most tumor cell lineages have undergone very large numbers of cell divisions, it should be considered a constant signature. If unaccounted for, it would likely find its way into other signatures, diminishing their accuracy. Matrix factorization methods have been developed that allow for fixing some elements of the solution (Kim et al. 2007, Limem et al. 2014, Gori and Baez-Ortega 2018), which enables fixing one or more signatures as a constant. While the concept of keeping some mutational signatures fixed has been used in a different context (Gori and Baez-Ortega 2018), to our knowledge separation of the normal mutational profile has not been proposed before.

(3) State-of-the-art NMF-based methods require the number of signatures as an input parameter but lack a principled basis for its selection. Discovering more signatures will always tend to reduce the residual error, i.e., fit the observed data better. However, the goal of signature discovery is not to fit the data as well as possible, but to identify signatures that truly reflect separate biological

processes. Currently, standard ways to choose the number of signatures are: (1) choosing a number such that more signatures would not significantly reduce residual error (Gehring et al. 2015); (2) choosing a number based on both minimizing residual error and maximizing reproducibility of signatures (Alexandrov et al. 2013); (3) calling signatures hierarchically on subsets of samples, adding more signatures in order to fit every sample (Nik-Zainal et al. 2016). The first two practices are ambiguous, while the third selects as many signatures as needed to improve fitting of the data, with little constraint to prevent overfitting. Overfitting can lead to many spurious signatures that actually represent noise, making it difficult to reliably attribute mutations in a sample to any one signature, leading to misinterpretation of the results and misleading conclusions. However, successful methods have been developed to choose the number of factors in NMF, including missing values imputation (the NNLM R package) and cross-validation (Mazumder et al. 2010). A recent signature discovery method, SignatureAnalyzer, uses automatic relevance determination, which starts with a high number of signatures and attempts to eliminate signatures of low relevance (Tan et al. 2013).

To overcome these drawbacks, we developed SparseSignatures (Figure 1A), a novel framework for mutational signature discovery. Like other NMF-based methods, SparseSignatures both identifies the signatures in a dataset of point mutations and calculates their exposure values (the number of mutations originating from each signature) in each patient.

Results

The SparseSignatures Algorithm. SparseSignatures is implemented in R and is available as a Bioconductor package at https://bioconductor.org/packages/release/bioc/html/SparseSignatures.html. Noteworthy innovations are:

1. It incorporates an explicit background model (Figure 1B), based on the human germline mutation spectrum (Rahbari et al. 2016), and validated in normal tissue samples (Supplementary Methods). Before using it, we made an empirical adjustment to CpG > TpG mutation rates (see Methods) of this background. This is because CpG > TpG mutations are frequently caused by cytosine deamination at sites of CpG methylation. Since the extent of CpG methylation can vary greatly in cancer cells, the rate of such mutations is not perfectly correlated with replication rates in tumors. SparseSignatures fixes the background signature and then discovers additional signatures representing cancer-specific mutagenic processes (including, usually, deamination of methylated cytosines). We note that our background signature is highly similar to 'Signature 5' in COSMIC, which has been found in cancer samples from numerous tissue sites as well as in normal somatic tissue (Alexandrov et al. 2015; Martincorena et al. 2018). This pattern of mutations is a signal of DNA replication error that occurs in all cells, and its clear presence across tissues supports our choice to incorporate a constant background model.
2. It uses LASSO regularization (Tibshirani 1996) to reduce noise in the signatures, except for the fixed background signature. The extent of regularization is controlled by a learned parameter, λ, for the entire signature matrix. We note that if the underlying signatures are very different in sparsity, this could result in a few individual signatures being too sparse or too dense if the value of λ is not ideal for them. However, we aim to improve the overall solution, and so our method chooses the best overall value of λ based on the complete dataset. It is also capable of choosing $\lambda=0$ (no LASSO penalty) if regularization does not in fact improve the solution.
3. It implements repeated bi-cross-validation (Owen et al. 2009) to select the best values for both the regularization parameter (λ) and the number of signatures (K). A randomly chosen subset of data points is held out and signatures are discovered based on the rest of the data. The values of the held-out data points are predicted based on the discovered signatures and their fitted exposure

values in each patient, and the mean squared error of the predictions is calculated. This procedure is performed for different values of K and λ, and the values that minimize the error in predicting held-out data points are chosen. The goal is to avoid overfitting, by ensuring that the discovered signatures not only fit the data used for discovery but also predict unseen values with high accuracy. In contrast to several previous methods, this provides a clear, unambiguous metric to choose the number of signatures.

SparseSignatures accurately detects signatures in simulated data. We compared SparseSignatures to two existing NMF-based methods for signature discovery, SigProfiler (Alexandrov et al. 2013) and SignatureAnalyzer (Tan et al. 2013). These two methods were the basis for a recent pan-cancer study (Alexandrov et al. 2020) resulting in 49 and 60 putative signatures. We generated 50 simulated datasets of 116 patients each with 4 underlying mutational signatures, based on curated WGS data from a cohort of Prostate cancer patients (see Methods). The underlying mutational signatures included a dense signature (COSMIC SBS3) as well as relatively sparse signatures (COSMIC SBS1, SBS18). We applied all three methods for signature discovery to this simulated dataset.

On simulated data, SparseSignatures is most effective at discovering the correct number of signatures (Figure 2A). SignatureAnalyzer consistently overfits the data, i.e., it overestimates the number of signatures and discovers an excessive number of sparse signatures that fit the data well but do not represent the actual underlying processes.

When comparing the overall residual error obtained by the three methods, SignatureAnalyzer fits the input matrix with the least residual error (Figure 2B-C). However, this is the result of overfitting as the method infers too many signatures. To provide a clearer measure, we assessed how well each method deciphers the input signatures by matching each of the input signatures to the most similar signature produced by the method, and assessing the residual error between these pairs of signatures. We did not

include the background signature in this comparison. Compared to other methods, SparseSignatures reconstructs the input signatures more accurately (Figure 2D). We also compared the original exposure values for each input signature to the exposure values produced by the method for the closest deciphered signature, and found that SparseSignatures shows lower error in reconstructing the original exposure values (Figure 2E).

Appropriate regularization of the signatures based on a learned parameter (λ) is one reason for the higher accuracy of our approach. The sparsity of signatures deciphered by SparseSignatures closely matches that of the input signatures (Figure 2F). In comparison, SigProfiler (which does not include any regularization term) tends to discover signatures with the addition of considerable noise, while SignatureAnalyzer produces excessively sparse signatures.

While SparseSignatures is the most accurate method at discovering the correct number of signatures, we also compared the performance of all three methods if the correct number of signatures is already known. When all three methods were given the correct number of signatures, SparseSignatures was still the most accurate at reconstructing the input data, signatures, and exposures (Supplementary Figure 1).

To provide additional validation of the robust performance of SparseSignatures, we performed three additional simulation experiments with different types of underlying signatures.

(1) We generated simulated data using randomly selected signatures from the COSMIC version 3 database.

(2) We generated simulated data using randomly selected signatures from the COSMIC version 3 database, limited to relatively dense signatures where >75% of the 96 mutation types contribute to the signature.

(3) We generated simulated data using randomly selected signatures from the COSMIC version 3 database, limited to relatively sparse signatures where <50% of the 96 mutation types contribute to the signature.

In all three additional simulations, we obtained similar results (Supplementary Figures 2-4). SignatureAnalyzer performs poorly at discovering the number of signatures; SparseSignatures and SigProfiler both perform better, frequently identifying the correct number of signatures or coming close. However, SparseSignatures, is more accurate than SigProfiler at reconstructing both the input signatures and exposures. Overall, SparseSignatures exceeds the performance of both other methods. . This shows the robust performance of SparseSignatures and its ability to accurately reconstruct input signatures with different characteristics.

SparseSignatures discovers well-differentiated signatures in pancreatic cancer data. We applied SparseSignatures to a dataset of patients affected by pancreatic cancer from PCAWG, including 147 curated whole genomes (Supplementary Table 1). Our goal was to discover mutational signatures that can be reconstructed with high accuracy and confidence. We therefore limited our analysis only to high-quality genomes with at least 1000 point mutations.

SparseSignatures discovered 8 signatures in addition to the background (Figure 3, Supplementary Tables 2 and 3) along with their exposure values for each patient (Supplementary Table 4). We named these discovered signatures in the format “PC-SS”, for “Pancreatic Cancer - SparseSignatures”. We compared these signatures to literature on known mutational mechanisms and to the signatures described in the COSMIC database. Remarkably, most of the signatures can be associated with a known mutational process (Table 1). For example, PC-SS1 is caused by deamination of methylated cytosine in CpG contexts, and PC-SS2 and PC-SS5 by APOBEC enzymes.

We ran both SigProfiler and SignatureAnalyzer on the same dataset for comparison. SigProfiler selected 9 as the optimal number of signatures (Supplementary Figure 5); SignatureAnalyzer selected 8 as the best number of signatures (Supplementary Figure 6).

Compared to the solution produced by SigProfiler, our signatures fit the observed data better (Table 2, Supplementary Table 5). While our solution and the solution produced by SigProfiler show similar accuracy at fitting the input data, our signatures are sparser, and show the lowest similarity between signatures, indicating that they are more clearly differentiated from each other. Moreover, our signatures show the lowest similarity between the background and the non-background signatures, suggesting that the other sets contain noise due to imperfect separation of the background signature (Table 2).

This is supported by visual inspection of the signatures predicted by the three methods. The signatures predicted by SigProfiler appear to contain visible background noise (Supplementary Figure 4). In addition, SPR7 seems to result from imperfect separation of one of the well-known APOBEC mutagenesis signatures (PC-SS4), while SPR4 shows a low level of contamination with the CpG deamination signature (PC-SS1). The signatures produced by SignatureAnalyzer appear to lack the low level of background noise throughout, but show similar imperfect separation of APOBEC signatures in SIA3 and SIA7 (Supplementary Figure 5).

Exposures predicted by SparseSignatures identify pancreatic cancer subtypes and correlate with clinical features. We next examined the exposure values produced by SparseSignatures for the background and 8 newly predicted signatures in pancreatic cancer samples. PC-SS1 (cytosine deamination at sites of CpG methylation) is the dominant signature, followed by PC-SS4 (background) and PC-SS7 (possibly reactive oxygen species) (Figure 4A).

We clustered all 147 tumors using CIMLR (Ramazzotti et al. 2018) based on these exposure values in order to identify subpopulations of tumors with similar mutagenic mechanisms. Using a bootstrap-based approach (Supplementary Methods, Wang et al. 2017, Wang et al. 2018b), we identified 10 clusters (Supplementary Figure 7, Supplementary Table 6) with different underlying exposures to the same set of 8 signatures (Figure 4B). C10 is high for PC-SSS3 (likely representing defective homologous recombination-based DNA damage repair). PC-SS4 is high in cluster 9, PC-SS7 is high in cluster 7, while cluster 8 seems to have high APOBEC signatures (PC-SS2 + PC-SS5).

The exposure-based clusters correlate with clinical features; cluster C1 with high PC-SS1 is enriched for females (Hypergeometric test p=0.0066), while cluster C10 has younger patients than the rest of the population (Wilcoxon test p=0.0333). Finally, patient relapse-free survival is significantly different between clusters, showing the potential clinical value of accurate signature discovery (Figure 4C). In contrast, the exposure values predicted by SignatureAnalyzer and SigProfiler are less effective at clustering patients into survival-associated subtypes (Supplementary Figure 8).

Discussion

SparseSignatures is a novel approach designed to discover the best number of clearly differentiated mutational signatures with minimal background noise, which have robust statistical support by repeated cross-validation on unseen data points and are not likely due to overfitting. Complementing its methodological innovations, SparseSignatures models a universal biological process in the form of a constant background signature. The biological motivation is the fact that all cells are subject to replication error, which is the result of misincorporation of nucleotides and subsequent failure of the proofreading mechanisms of the DNA polymerases. Other processes such as transcription-coupled repair also contribute to the 'normal' mutation burden (Green et al. 2013). Although cell culture mutation spectra have been estimated (Milholland et al. 2017), we chose to base our background on the human germline

signature, which is currently the most robust estimate of a non-cancer *in vivo* mutation spectrum. Our result that the background signature is the most dominant signature overall provides empirical evidence that it is in fact prevalent in our data.

Multiple experiments on simulated data show that SparseSignatures outperforms current state-of-the-art methods. It provides the most accurate and least ambiguous estimation of the number of signatures, and reconstructs the original signatures and exposures most accurately. In comparison, other methods tend to discover too many signatures or retain noise in the discovered signatures.

Using SparseSignatures on whole genome sequences from 147 pancreatic tumors, we obtain 8 signatures in addition to the background. Compared to other methods, we successfully obtain a good fit to the observed data, while at the same time obtaining signatures that are sparse, differentiated, have reduced noise, and are attributable to known biological processes while at the same time preventing overfitting.

We have also applied our method to a larger breast cancer dataset (Lal et al. 2019). On this dataset, we were able to discover additional signatures (such as a signature associated with BRCA1 and BRCA2 mutations), some of which contributed only to a small subset of samples. We anticipate that the availability of larger datasets comprising curated, uniformly processed whole genome sequences may allow us to validate those signatures and discover new ones.

Our method supports the discovery of sparse signatures by applying a LASSO penalty to the signatures matrix. Currently no penalty is applied to the exposure matrix which consequently is non-sparse. It is also reasonable to believe that only a limited number of mutational processes will be active in each patient, and therefore even better results may be obtained by supporting increased sparsity in the exposures matrix as well. While this option is not currently available due to its high computational cost, we aim to support it in a future version. Future work could also be directed at incorporating indels and doublet base substitutions (Alexandrov et al. 2020), especially when larger datasets become available to support analyses of these rarer events.

In conclusion, the small number of highly specific, differentiated signatures discovered by SparseSignatures leads us to predict that whole genome sequencing of individual cancers and their classification on the basis of signatures, including the background, may become much more easily interpretable and possibly useful in a clinical context. For example, strong contribution of CpG methylation versus background in a patient suggests that methylation changes have been more important for the growth of the cancer and that overall cellular turnover (associated with background) may have been modest, suggesting that DNA replication inhibitors may be less effective than gene regulatory therapy for such patients.

We suggest that future work be directed at greater numbers of patients for whole genome sequencing and the simultaneous collection of other omic data to connect mutagenesis with molecular phenotype and eventually mechanistic cause.

Methods

Mathematical Framework for Mutational Signature Discovery. The mathematical framework developed for signature extraction (Alexandrov et al. 2013) is as follows. First, all point mutations are classified into 6 groups (C>A, C>G, C>T, T>A, T>C, T>G; the original pyrimidine base is listed first). Then, these are subdivided into 16 x 6 = 96 categories based on the 16 possible combinations of 5' and 3' flanking bases. Each tumor sample is described by the count of mutations in each of the 96 categories. This forms a count matrix M, where the rows are the tumor samples and the columns are the 96 categories.

Signature extraction aims to decompose M into the multiplication of two low-rank matrices: the exposure matrix α and the signature matrix β.

$$M \approx \alpha\beta \quad \text{(Equation 1)}$$

Here, α is the exposure matrix with one row per tumor and K columns, and β is the signature matrix with K rows and 96 columns. K is the number of signatures. Each row of β represents a signature, and each row of α represents the exposure of a single tumor to all K signatures, i.e. the number of mutations contributed by each signature to that tumor. In NMF, this equation is solved for α and β by minimizing the squared residual error (some methods use Kullback–Leibler divergence instead) while constraining all elements of α and β to be non-negative.

$$min\ ||M - \alpha\beta||^2_F \ \ subject\ to\ \alpha \geq 0,\ \beta \geq 0$$

Improvements to the NMF framework in SparseSignatures. In SparseSignatures, we incorporate a background signature by modifying Equation (1) as follows:

$$M \approx \alpha_0\beta_0 + \alpha\beta \quad \text{(Equation 2)}$$

Here, β_0 is the known 'background' signature of point mutations caused by replication errors during cell division, and α_0 is the vector of exposures of all tumors to that signature. The dimensions of α_0 are (number of tumors x 1) and the dimensions of β_0 are 1 x 96.

To enforce sparsity in the discovered signatures, we use the LASSO (Tibshirani 1996). This is done by adding an additional regularization term to the cost function to be minimized:

$$min\ ||M - (\alpha_0\beta_0 + \alpha\beta)||_F^2 + \lambda||\beta||_1\ subject\ to\ \alpha \geq 0,\ \beta \geq 0,\ \alpha_0 \geq 0$$

The parameter λ controls the extent to which sparsity is encouraged in the signature matrix β. If the value of λ is set too low, it is ineffective, whereas if it is set too high, the signatures are forced to be too sparse and no longer accurately fit the data.

It should be noted that unlike the standard LASSO, the objective function we minimize here is non-convex. But it is bi-convex (convex in α with β fixed and vice-versa). Hence the alternating algorithm described below is natural and yields good solutions. A standard issue with all NMF algorithms is non-identifiability: if we scale β by c and α by $1/c$, the product αβ remains unchanged. One can change

the relative magnitudes of α and β at convergence by changing their relative magnitudes at initialization. To remove this ambiguity, we initialize β so that each row (signature) sums to 1. The choice of 1 is not important: if we had instead initialized β so that each row sums to c, the signatures we obtain at algorithm convergence would be equivalent (up to proportionality) to those obtained by initializing β with all rows summing to 1 and λ set to λ/c.

Implementation of SparseSignatures. SparseSignatures discovers mutational signatures by following the steps below.

Step 1: Build the Count Matrix M by counting the number of mutations of each of the 96 categories in each sample.

Step 2: Remove samples with less than a minimum number of mutations. In the analysis described in this paper, we have used a minimum number of 1000 mutations per tumor genome.

Step 3: Choose a range of values to test for K (number of signatures) and λ (level of sparsity).

Step 4: For each value of K in the chosen range, obtain a set of K initial signatures using repeated NMF (Brunet et al. 2004) to obtain a more robust estimation. This is an initial value for the matrix β. We use these NMF results as a starting point (although other starting points such as randomly generated signatures may also be chosen) and further refine the signatures. In practice, the final discovered signatures are often very different from those produced by the initial NMF.

Step 5: For each pair of parameter values (K and λ), perform cross-validation as follows (Mazumder et al. 2010):

5a. Randomly select a given percentage of cells from M. Based on simulations (Supplementary Methods, Supplementary Table 7), we currently use 1% of the points in the dataset for cross-validation; however, the method appears robust to large variations in this value.

5b. Replace the values in those cells with 0.

5c. Consider the NMF results for the chosen value of K as an initial value of β. Add the background signature (β_0). Then use an iterative approach to discover signatures with sparsity. Each iteration involves two steps:

5c(i). While keeping fixed the values of β_0 and β, fit α_0 and α by minimizing:

$$min\ ||M - (\alpha_0\beta_0 + \alpha\beta)||_F^2\ subject\ to\ \alpha \geq 0,\ \alpha_0 \geq 0$$

5c(ii). While keeping fixed the values of β_0, α_0 and α, fit β by minimizing:

$$min\ ||M - (\alpha_0\beta_0 + \alpha\beta)||_F^2 + \lambda||\beta||_1\ subject\ to\ \beta \geq 0$$

These steps are repeated for a number of iterations (set to 20 by default; in all our experiments we found that this was sufficient to reach convergence).

5d. Use the obtained signatures to predict the values for the cells that were set to 0 (we do this by calculating the matrix $\alpha_0\beta_0 + \alpha\beta$ and taking the entries corresponding to the cross-validation cells). Then replace the values in these cells with the predicted values and repeat step 5c. We repeat step 5c a number of times (set to 5 by default), each time discovering signatures and then replacing the values of the cross-validation cells by the predicted values. After each iteration, the predictions improve, as the algorithm converges, making the mean squared errors used in the next step more stable.

5e. At the last iteration of step 5d, measure the mean squared error (MSE) of the prediction.

5f. Repeat the entire cross-validation procedure (steps 5a-5d) a number of times (set to 10 by default) and calculate the MSE for all cross-validations. Since we randomly select a different set of cells for cross-validation each time, this allows us to obtain a robust measure of MSE.

Step 6: Choose the values of K and λ that correspond to the lowest MSE in most of the cross-validations.

Step 7: Using the selected values for K and λ, repeat sparse signature discovery (step 5c) on the complete matrix M (without replacing any cells with 0). This generates the final values of α_0, α and β.

Background signature. We used the germline mutation spectrum calculated by (Rahbari et al 2016) as our background signature. To validate this, we independently calculated the germline mutational spectrum using whole-genome sequencing data from normal tissue samples (see Supplementary Methods for details), and the spectrum thus obtained had a high cosine similarity of 0.98 with that calculated by (Rahbari et al 2016). We then adjusted the rates of ACG>ATG, CCG>CTG, GCG>GTG and TCG>TTG mutations to be equal to the rates of ACA>ATA, CCA>CTA, GCA>GTA and TCA>TTA mutations respectively, in order to separate the effects of DNA methylation from the background signature. We also compared our background signature with the COSMIC signature 5, which has been associated with aging, and found a high cosine similarity of 0.93.

Definition of the λ parameter. This parameter tunes the desired level of regularization to be obtained by LASSO. For any analysis by LASSO, one can compute a maximal value of the LASSO penalty after which all the coefficients of the regression get shrunk to zero (Friedman et al. 2010). As this maximal value can vary depending on the problem, our λ parameter represents the fraction of the actual maximal value to be used. Values closer to 1 result in higher regularization.

Simulations. We generated 4 experiments all including 50 simulated datasets of 4 signatures. In experiment 1, we used real data to perform simulations and specifically we considered 116 curated WGS data of prostate cancer samples obtained from PCAWG (https://dcc.icgc.org/pcawg) with at least 1000 mutations, and selected a set of 4 signatures from COSMIC known to be active in prostate (Alexandrov et al. 2020); we then used deconstructSigs (Rosenthal et al. 2016) to fit such signatures on the data and generate their assignments to samples. Furthermore, we performed three additional experiments where we randomly selected signatures from COSMIC, considering all of them (experiment 2) and the subset of

dense (experiment 3) or sparse (experiment 4) signatures; we then generated random assignments of such signatures to samples, for a total of 100 samples per experiment.

We ran three methods for *de novo* signature discovery (SparseSignatures, SignatureAnalyzer, and SigProfiler) on each of the 50 datasets and evaluated their performance. These methods were executed with the configurations suggested by the authors in the respective manuscripts. Specifically, SignatureAnalyzer was performed 10 times and the solution with best posterior was chosen; SigProfiler pipeline was performed 10 times with 100 iterations each. Details are provided as Supplementary Methods. To evaluate the accuracy with which discovered signatures reconstructed the original signatures, we matched each input signature to its closest discovered signature and evaluated the match by mean squared error. We then also measured the mean squared error between the exposure values of the input signature and the discovered exposure values for its most similar discovered signature. Further details are given in the Supplementary Methods (Experiments 1, 2, 3 and 4).

Pancreatic cancer dataset. We obtained a dataset of point mutations from ICGC (see Supplementary Table 1 for the full list of samples). We selected only whole-genome sequencing data and removed samples with less than 1000 point mutations. After this preprocessing, a total of 147 samples remained.

Software. The experiments carried out in this paper were performed using the SparseSignatures v1.0.1 R package and R version 3.4.3. The software is available for download on Bioconductor at https://bioconductor.org/packages/release/bioc/html/SparseSignatures.html. This package in its current version makes use of external R packages NMF v0.21.0 (Gaujoux and Seoighe 2010), nnls v1.4 and nnlasso v0.3.

Data Access

The whole-genome sequencing data used in this study is publicly available and was downloaded from https://dcc.icgc.org/search.

Acknowledgments

This work was supported by an R01 grant to A.S. (NIH/NCI) and gift funding from the BRCA Foundation. A.L. was supported by a Young Investigator Award from the BRCA Foundation. The results published here are based in part upon data generated by the Pan-Cancer Analysis of Whole Genomes (PCAWG) Research Network (https://dcc.icgc.org/pcawg).

Disclosure declaration

The authors declare that there are no conflicts of interest.

References

1. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., & Stratton, M. R. (2013). Deciphering signatures of mutational processes operative in human cancer. Cell reports, 3(1), 246-259.

2. Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., & Stratton, M. R. (2015). Clock-like mutational processes in human somatic cells. Nature genetics, 47(12), 1402.

3. Alexandrov, L. B., Ju, Y. S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., ... & Campbell, P. J. (2016). Mutational signatures associated with tobacco smoking in human cancer. Science, 354(6312), 618-622.

4. Alexandrov, Ludmil B., et al. "The repertoire of mutational signatures in human cancer." Nature 578.7793 (2020): 94-101.

5. Bolli, N., Avet-Loiseau, H., Wedge, D. C., Van Loo, P., Alexandrov, L. B., Martincorena, I., ... & Hinton, J. W. (2014). Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. Nature communications, 5, 2997.

6. Brunet, J. P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the national academy of sciences, 101(12), 4164-4169.

7. Covington, K., Shinbrot, E., & Wheeler, D. A. (2016). Mutation signatures reveal biological processes in human cancer. bioRxiv, 036541.

8. Fischer, A., Illingworth, C. J., Campbell, P. J., & Mustonen, V. (2013). EMu: probabilistic inference of mutational processes and their localization in the cancer genome. Genome biology, 14(4), R39.

9. Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1), 1.

10. Gaujoux, R., & Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. BMC bioinformatics, 11(1), 367.

11. Gehring, J. S., Fischer, B., Lawrence, M., & Huber, W. (2015). SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics, 31(22), 3673-3675.

12. Goncearenco, A., Rager, S. L., Li, M., Sang, Q. X., Rogozin, I. B., & Panchenko, A. R. (2017). Exploring background mutational processes to decipher cancer genetic heterogeneity. Nucleic acids research, 45(W1), W514-W522.

13. Gori, K., & Baez-Ortega, A. (2018). sigfit: flexible Bayesian inference of mutational signatures. bioRxiv, 372896.

14. Green, P., Ewing, B., Miller, W., Thomas, P. J., Green, E. D., & NISC Comparative Sequencing Program. (2003). Transcription-associated mutational asymmetry in mammalian evolution. Nature genetics, 33(4), 514.

15. Helleday, T., Eshtad, S., & Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. Nature Reviews Genetics, 15(9), 585.

16. Kim, H., & Park, H. (2007). Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics, 23(12), 1495-1502.

17. Lal, A., Ramazzotti, D., Weng, Z., Liu, K., Ford, J. M., & Sidow, A. (2019). Comprehensive genomic characterization of breast tumors with BRCA1 and BRCA2 mutations. BMC Medical Genomics, 12(1), 84.

18. Ledford, H. (2017). DNA typos to blame for most cancer mutations. Nature News.

19. Limem, A., Delmaire, G., Puigt, M., Roussel, G., & Courcot, D. (2014). Non-negative matrix factorization under equality constraints—a study of industrial source identification. Applied Numerical Mathematics, 85, 1-15.

20. Martincorena, I., Fowler, J. C., Wabik, A., Lawson, A. R., Abascal, F., Hall, M. W., ... & Fitzgerald, R. C. (2018). Somatic mutant clones colonize the human esophagus with age. Science, eaau3879.

21. Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. Journal of machine learning research, 11(Aug), 2287-2322.

22. Milholland, B., Dong, X., Zhang, L., Hao, X., Suh, Y., & Vijg, J. (2017). Differences between germline and somatic mutation rates in humans and mice. Nature communications, 8, 15183.

23. Nik-Zainal, S., Kucab, J. E., Morganella, S., Glodzik, D., Alexandrov, L. B., Arlt, V. M., ... & Phillips, D. H. (2015). The genome as a record of environmental exposure. Mutagenesis, 30(6), 763-770.

24. Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., ... & Van Loo, P. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature, 534(7605), 47.

25. Owen, A. B., & Perry, P. O. (2009). Bi-cross-validation of the SVD and the nonnegative matrix factorization. The annals of applied statistics, 3(2), 564-594.

26. Pascual-Montano, A., Carazo, J. M., Kochi, K., Lehmann, D., & Pascual-Marqui, R. D. (2006). Nonsmooth nonnegative matrix factorization (nsNMF). IEEE transactions on pattern analysis and machine intelligence, 28(3), 403-415.

27. Rahbari, R., Wuster, A., Lindsay, S. J., Hardwick, R. J., Alexandrov, L. B., Al Turki, S., ... & Stratton, M. R. (2016). Timing, rates and spectra of human germline mutation. Nature genetics, 48(2), 126.

28. Ramazzotti, D., Lal, A., Wang, B., Batzoglou, S., & Sidow, A. (2018). Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. Nature communications, 9(1), 4453.

29. Rosenthal, Rachel, et al. "DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution." Genome biology 17.1 (2016): 1-11.

30. Schulze, K., Imbeaud, S., Letouzé, E., Alexandrov, L. B., Calderaro, J., Rebouissou, S., ... & Calatayud, A. L. (2015). Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nature genetics, 47(5), 505.

31. Shiraishi, Y., Tremmel, G., Miyano, S., & Stephens, M. (2015). A simple model-based approach to inferring and visualizing cancer mutation signatures. PLoS genetics, 11(12), e1005657.

32. Tan, V. Y., & Fevotte, C. (2013). Automatic relevance determination in nonnegative matrix factorization with the/spl beta/-divergence. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(7), 1592-1605.

33. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

34. Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. science, 339(6127), 1546-1558.

35. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., & Batzoglou, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nature methods, 14(4), 414.

36. Wang, S., Jia, M., He, Z., & Liu, X. S. (2018). APOBEC3B and APOBEC mutational signature as potential predictive markers for immunotherapy response in non-small cell lung cancer. Oncogene.

37. Wang, B., Ramazzotti, D., De Sano, L., Zhu, J., Pierson, E., & Batzoglou, S. (2018). SIMLR: A Tool for Large‐Scale Genomic Analyses by Multi‐Kernel Learning. Proteomics, 18(2), 1700232.

Tables

Signature	Proposed etiology	Basis for proposed etiology
PC-SS 1	Cytosine methylation / deamination	Experimental evidence (Helleday et al. 2014)
PC-SS 2	APOBEC dysregulation	Experimental evidence (Helleday et al. 2014)
PC-SS 3	Defective homologous recombination-based DNA damage repair	Hypothesised (Alexandrov et al. 2020)
PC-SS 4 (Background)	DNA replication error	Experimental evidence (Rahbari et al. 2016)
PC-SS 5	APOBEC dysregulation	Experimental evidence (Helleday et al. 2014)
PC-SS 6	Unknown	N/A
PC-SS 7	Damage by reactive oxygen species	Hypothesised (Alexandrov et al. 2020)
PC-SS 8	Defective DNA mismatch repair	Hypothesised (Alexandrov et al. 2020)
PC-SS 9	Possible sequencing artefact	Hypothesised (Alexandrov et al. 2020)

Table 1. 9 signatures (including background, Signature 4) discovered by SparseSignatures and their proposed etiology.

Source	Number of signatures	MSE	Sparsity (signatures) (Fraction of cells with value < 0.001)	Average similarity between signatures	Average similarity between background and non-background signatures	Median per-patient correlation between observed and predicted counts
SparseSignatures	9	1189.521	0.37	0.19	0.37	0.99
SigProfiler	9	1003.428	0.15	0.37	0.51	0.98
SignatureAnalyzer	8	52606.28	0.32	0.24	0.45	0.99

Table 2. Comparison of signatures predicted by three methods on 147 pancreatic cancer whole genomes.

Figure Legends

Figure 1. A) Schematic of the SparseSignatures method. N represents the number of tumors in the dataset, K the number of signatures. B) Background signature based on the human germline mutation spectrum. Vertical bars represent the probability of mutation in each of 96 categories. These are based on 6 possible mutation types (upper gray labels) and 16 possible combinations of 5' and 3' flanking bases (x-axis labels).

Figure 2. Comparison between SparseSignatures and other methods on simulated data. A) Bar and line plot showing, for each method, the number of times it selected each value of K (number of signatures). The x-axis shows values of K and the y-axis shows the number of times each value was selected. Each method was run on 50 simulated datasets. In all cases, the correct value of K was 4. B) Box plots showing the fraction of variance in the count matrix explained by the solutions produced by each method, over 50 simulations. C) Box plots showing the residual error for the solutions produced by each method, over 50 simulations. Residual error was measured as the mean squared error (MSE) in reconstructing the original count matrix. D) Box plots showing the mean squared error in reconstructing the 3 non-background input signatures, over 50 simulations. E) Box plots showing the mean squared error in reconstructing the exposure values for the 3 non-background input signatures, over 50 simulations. F) Box plots showing the sparsity of the signatures produced by each method, over 50 simulations. The gray box shows the sparsity of the input signatures. Sparsity was measured as the fraction of cells in the signature matrix whose value is $<10^{-3}$.

Figure 3. The 9 mutational signatures obtained by applying SparseSignatures to a dataset of 147 pancreatic tumors. We report the number and correlation of the most similar (correlation higher than 0.70) corresponding signature from COSMIC (https://cancer.sanger.ac.uk/cosmic/signatures).

Figure 4. (A) Fitted values for exposure to each of the 9 signatures obtained by SparseSignatures for the 147 pancreatic tumors. Each panel shows boxplots representing the fraction of mutations per tumor (on the y-axis) contributed by the given signature (on the x-axis). (B) Clustering of patients based on their exposure values. Boxplots show the fraction of mutations per tumor contributed by each signature (x-axis) to each of 10 clusters. (C) Relapse-free survival analysis of patients belonging to the 10 clusters.

Supplementary Material Legends

Supplementary Figure 1. Comparison between SparseSignatures and other methods on simulated data when the correct number of signatures is known. A) Box plots showing the residual error for the solutions produced by each method, over 50 simulations. Residual error was measured as the mean squared error (MSE) in reconstructing the original count matrix. B) Box plots showing the mean squared error in reconstructing the 3 non-background input signatures, over 50 simulations. C) Box plots showing the mean squared error in reconstructing the exposure values for the 3 non-background input signatures, over 50 simulations. D) Box plots showing the sparsity of the signatures produced by each method, over 50 simulations. The gray box shows the sparsity of the input signatures. Sparsity was measured as the fraction of cells in the signature matrix whose value is $<10^{-3}$.

Supplementary Figure 2. Comparison between SparseSignatures and other methods on simulated data generated from 4 randomly selected COSMIC signatures. A) Box plots showing the residual error for the solutions produced by each method, over 50 simulations. Residual error was measured as the mean squared error (MSE) in reconstructing the original count matrix. B) Box plots showing the fraction of variance in the count matrix explained by the solutions produced by each method, over 50 simulations. C) Box plots showing the mean squared error in reconstructing the 3 non-background input signatures, over 50 simulations. D) Box plots showing the mean squared error in reconstructing the exposure values for the 3 non-background input signatures, over 50 simulations. E) Box plots showing the sparsity of the signatures produced by each method, over 50 simulations. The gray box shows the sparsity of the input signatures. Sparsity was measured as the fraction of cells in the signature matrix whose value is $<10^{-3}$.

Supplementary Figure 3. Comparison between SparseSignatures and other methods on simulated data generated from 4 randomly selected dense COSMIC signatures. A) Box plots showing the residual error for the solutions produced by each method, over 50 simulations. Residual error was measured as the mean squared error (MSE) in reconstructing the original count matrix. B) Box plots showing the fraction of variance in the count matrix explained by the solutions produced by each method, over 50 simulations. C) Box plots showing the mean squared error in reconstructing the 3 non-background input signatures, over 50 simulations. D) Box plots showing the mean squared error in reconstructing the exposure values for the 3 non-background input signatures, over 50 simulations. E) Box plots showing the sparsity of the signatures produced by each method, over 50 simulations. The gray box shows the sparsity of the input signatures. Sparsity was measured as the fraction of cells in the signature matrix whose value is $<10^{-3}$.

Supplementary Figure 4. Comparison between SparseSignatures and other methods on simulated data generated from 4 randomly selected sparse COSMIC signatures. A) Box plots showing the residual error for the solutions produced by each method, over 50 simulations. Residual error was measured as the mean squared error (MSE) in reconstructing the original count matrix. B) Box plots showing the fraction of variance in the count matrix explained by the solutions produced by each method, over 50 simulations. C) Box plots showing the mean squared error in reconstructing the 3 non-background input signatures, over 50 simulations. D) Box plots showing the mean squared error in reconstructing the exposure values for the 3 non-background input signatures, over 50 simulations. E) Box plots showing the sparsity of the signatures produced by each method, over 50 simulations. The gray box shows the sparsity of the input signatures. Sparsity was measured as the fraction of cells in the signature matrix whose value is $<10^{-3}$.

Supplementary Figure 5. 9 signatures predicted by SigProfiler on 147 pancreatic tumors.

Supplementary Figure 6. 8 signatures predicted by SignatureAnalyzer on 147 pancreatic tumors.

Supplementary Figure 7. (A) CIMLR bootstrap (B) CIMLR number of clusters for SparseSignatures.

Supplementary Figure 8. (A) Survival for CIMLR clusters on SigProfiler. (B) Survival for CIMLR clusters on SignatureAnalyzer.

Supplementary Table 1. List of 147 Pancreatic cancer samples used for signature discovery.

Supplementary Table 2. Results of cross-validation to choose the best values of K and λ on pancreatic cancer data, using 1% of the cells in the matrix for cross-validation. We tested values of K ranging from 2 to 16 and values of lambda of 0.01, 0.025, 0.05, 0.075 and 0.1. Cross-validation was repeated 500 times with 5 restarts each. The entries in the table represent the median mean square error (MSE) in fitting the unseen data points across the 500 repetitions.

Supplementary Table 3. 9 signatures (including the background signature) discovered by applying SparseSignatures to pancreatic cancer data.

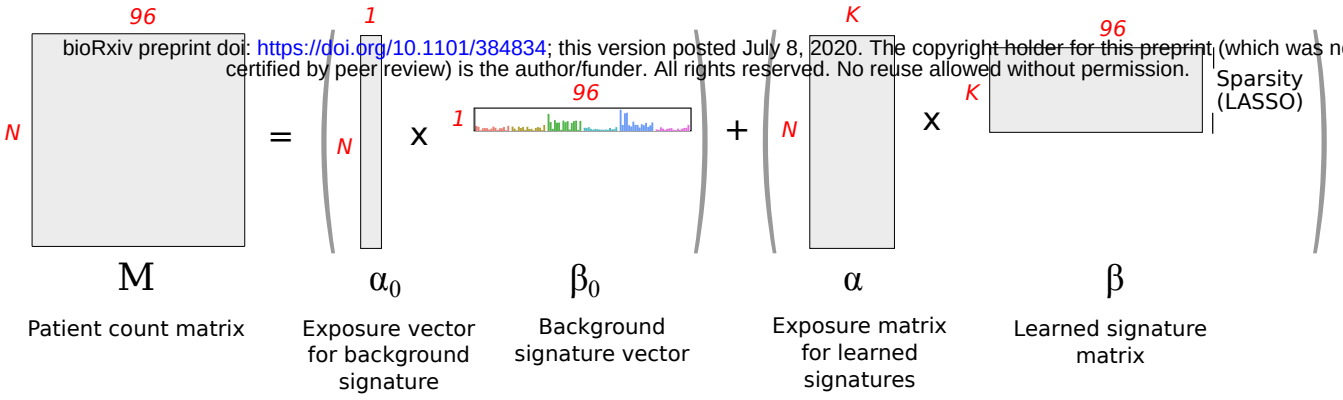
Supplementary Table 4. Fitted values for exposure to each of the 9 signatures (including the background signature) discovered by applying SparseSignatures to pancreatic cancer data, of each of the 147 whole genomes in the dataset.

Supplementary Table 5. Mean correlation of observed and predicted counts for each patient.

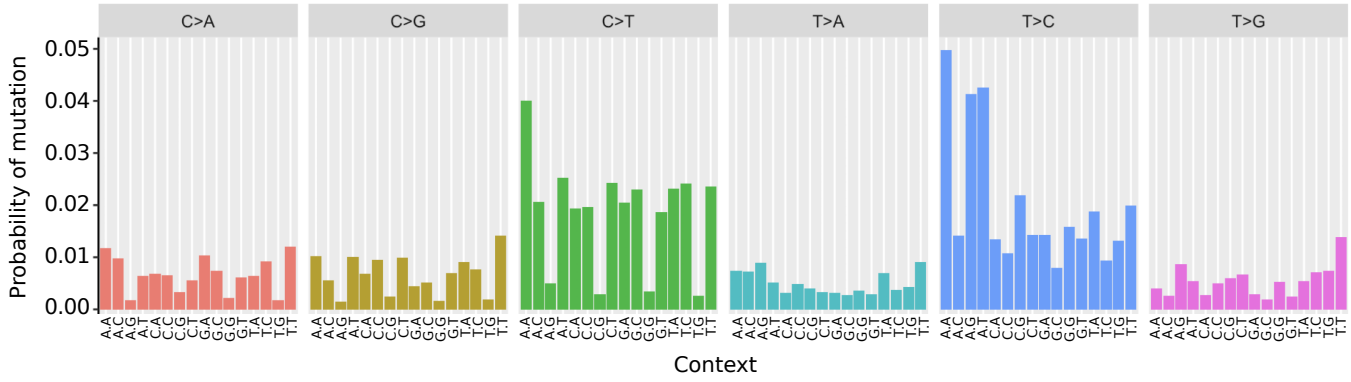
Supplementary Table 6. Cluster assignments generated by CIMLR for each sample.

Supplementary Table 7. Results of cross-validation to choose the best values of K on simulated data, using 0.1%, 1%, and 10% of the cells in the matrix M for cross-validation. Cross-validation was repeated 100 times for each percentage of cells. The entries in the table represent the median mean square error (MSE) in fitting the unseen data points across the 100 repetitions.

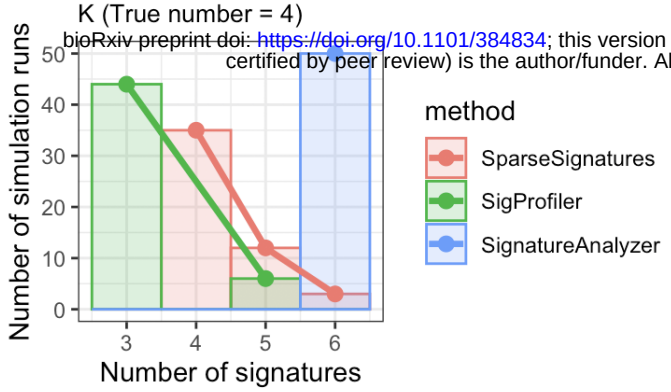
A



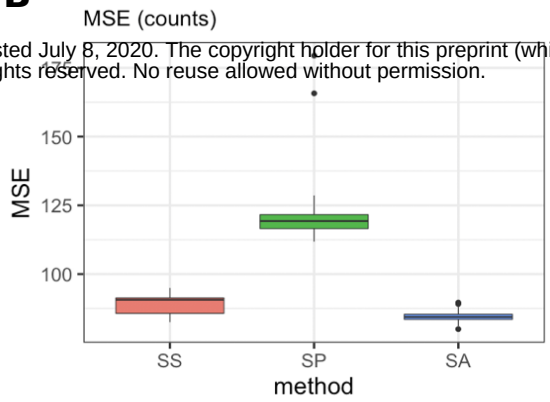
B



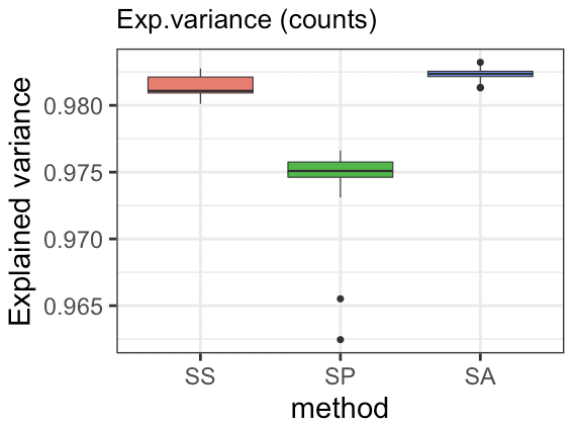
A



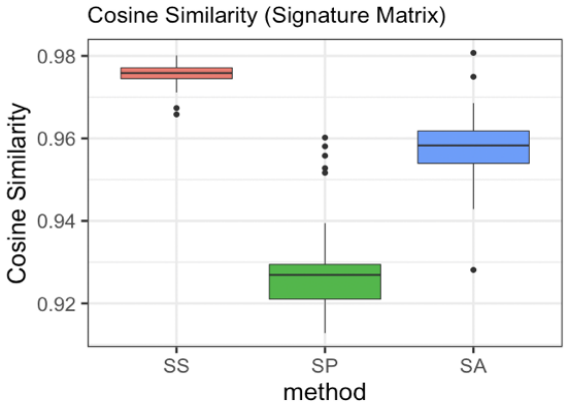
B



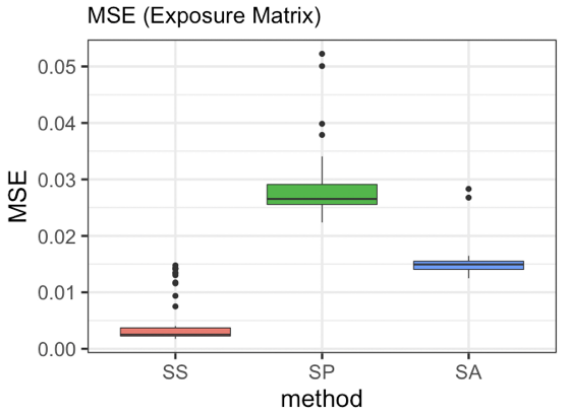
C



D



E



F