Mutational Signature Refitting on Sparse Pan-Cancer Data

Gal Gilad
Blavatnik School of Computer Science and AI, Tel Aviv University, Israel

Teresa M. Przytycka
National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD, USA

Roded Sharan[1]
Blavatnik School of Computer Science and AI, Tel Aviv University, Israel

Abstract

Mutational processes shape cancer genomes, leaving characteristic marks that are termed signatures. The level of activity of each such process, or its signature exposure, provides important information on the disease, improving patient stratification and the prediction of drug response. Thus, there is growing interest in developing refitting methods that decipher those exposures. Previous work in this domain was unsupervised in nature, employing algebraic decomposition and probabilistic inference methods. Here we provide a supervised approach to the problem of signature refitting and show its superiority over current methods. Our method, SuRe, leverages a neural network model to capture correlations between signature exposures in real data. We show that SuRe outperforms previous methods on sparse mutation data from tumor type specific data sets, as well as pan-cancer data sets, with an increasing advantage as the data become sparser. We further demonstrate its utility in clinical settings.

2012 ACM Subject Classification Applied computing → Bioinformatics

Keywords and phrases mutational signatures, signature refitting, cancer genomics, genomic data analysis, somatic mutations

Digital Object Identifier 10.4230/LIPIcs.WABI.2025.11

Supplementary Material *Dataset (Synthetic dataset)*: `https://doi.org/10.6084/m9.figshare.20409430.v1` [11]
Dataset (WGS mutation counts and exposures): `https://doi.org/10.7303/syn11726601` [2]
Software: `https://github.com/GalGilad/SuRe`
Dataset (WGS triple negative breast cancers): `https://doi.org/10.17632/2mn4ctdpxp.1` [19]

Funding This research was supported in part by a grant from the US-Israel Binational Science Foundation and in part by the Intramural Research Program of the National Library of Medicine, NIH, USA.

1 Introduction

The genomes of cancer cells accumulate somatic mutations throughout their developmental history. These mutations are the result of environmental and endogenous mutational processes that are active in each cell. The activity, or *exposure*, of these mutational processes in a genome can be revealed by the characteristic patterns of mutations they leave, termed *mutational signatures*. The discovery of such signatures is most commonly performed via non-negative matrix factorization (NMF); dozens of mutational signatures have been identified to date in thousands of sequenced genomes and exomes of diverse cancer types [2, 6]. These signatures are cataloged in the COSMIC database [20].

[1] To whom correspondence should be addressed

25th International Conference on Algorithms for Bioinformatics (WABI 2025).
Editors: Broňa Brejová and Rob Patro; Article No. 11; pp. 11:1–11:23
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Building on the reconstructed signatures, one of the main goals of mutational signature analysis is to infer the exposure of relevant signatures in a given sample from the mutations it harbors. This task is called *signature refitting*. Perhaps the most popular method in mutational signature analysis for inferring exposures from rich mutation data is non-negative least squares (NNLS) [9]. Typically, this method is used to find exposures that minimize the Kullback-Leibler divergence between the mutation counts and the product of the signatures and the exposures. Another method that can be used to refit mutational signatures is Mix [17], which was developed in order to cope with the problem of sparse data, typical in data derived from targeted sequencing panels. Mix simultaneously clusters the samples and learns signature exposures per cluster rather than per sample based on a probabilistic mixture model. Additional methods include: deconstructSigs [16], a heuristic based on the iterative application of the multiple linear regression and exclusion of signatures with low relative exposures; SigLASSO [13], which seeks to produce sparse, high-confidence solutions by jointly optimizing L1 regularized signature refitting and mutation sampling likelihood; SignatureEstimation [10] that evaluates the stability and confidence of signature activity levels by applying perturbations to the mutation count inputs; SigNet [18], which trains artificial neural networks based on labeled datasets to learn the prior frequencies of signatures and their correlations in whole-exome data; SigProfilerAssignment (also called SigProfilerAttribution) [2, 11], which takes into account previous knowledge about the signatures and their biological role, while iteratively adding and removing signatures to the optimization based on the true count reconstruction similarity; and MuSiCal [12], which applies a likelihood-based sparse NNLS approach.

Here, we take a different, supervised approach to mutational signature refitting, focusing on sparse mutations data (5 mutations on average when considering panel sequencing data). We train a neural network model that learns correlations between signature activity levels in the input data, in order to improve exposure prediction. We show that our model outperforms previous methods in the task of mutational signature refitting for sparse data, suggesting that it successfully captures such correlations within the data. Additionally, we demonstrate our model's ability to handle pan-cancer input data and exemplify its utility in a clinical setting.

2 Methods

2.1 Preliminaries

We focus on single base substitutions, the most common mutation type. A *mutation category* denotes the substitution that occurred and the flanking nucleotides. In the standard categorization there are 96 mutation categories, spanning six substitution options and the two flanking nucleotides, one on each side [3].

A *mutational signature* is a probability vector over mutation categories. Its *exposure* in a specific sample denotes the number of mutations in that sample that were caused by the signature. If we normalize an exposure by the total number of mutations in the sample we obtain a *relative exposure*. A *refitting* algorithm receives as input mutation data in a collection of samples and a set of known signatures. Its goal is to infer the exposure of each signature in each input sample. In a *de novo* setting the signatures are unknown and have to be inferred as well.

2.2 Mutation data

All main data sets were downloaded from two sources: i) previous analysis of COSMIC mutation data, based on a legacy version of SigProfiler, that was published in [2], and ii) re-analysis of the data using a newer version of the software, published in [11]. Specifically, for (i), we downloaded all available whole genome sequencing mutation counts (aka catalogs) in `https://www.synapse.org/#!Synapse:syn11726616`, as well as the corresponding patients' exposures to mutational signatures from `https://www.synapse.org/#!Synapse:syn11804040`. For (ii), the WGS data was downloaded from `https://doi.org/10.6084/m9.figshare.20409430`, along with synthetically generated samples that were previously used in [11] and [7] to benchmark extraction and assignment of mutational signatures.

We separated the mutation data into three sets: i) breast cancer data set consisting of 679 patients (195 PCAWG patients and 484 patients from the extended cohort); ii) pan-cancer data set consisting of 4,568 patients (2,703 from PCAWG and 1,865 from the extended cohort); and iii) synthetic data set of 2,700 simulated samples, comprising 300 samples from each of nine distinct cancer types, featuring single base substitution spectra matching those observed in PCAWG. In total, the breast cancer and pan-cancer data sets consist of a total of 4,427,411 and 72,222,147 mutations, respectively.

Our analyses are based on the COSMIC v3.3 mutational signatures for the GRCh37 human reference genome. According to the COSMIC analysis there are 20 active SBS (single base substitutions) signatures in the breast cancer data set, while there are 23 active SBS signatures according to the re-analysis.

In addition to these data sets, we have collected whole genome mutation data of triple negative breast cancer patients from Staaf et al. [19], along with their HRD status labels, predicted by HRDetect [5]. These labels are classified into three groups that represent the probability of HRD: high (score above 0.7), intermediate (0.2 to 0.7), and low (below 0.2). In total, 139 patients are predicted as " high" , while 13 and 85 are predicted as " intermediate" and " low" , respectively. We downsampled the mutation data using the MSK-IMPACT [4] panel and binarized the labels by excluding the 13 " intermediate" labeled samples, leaving 224 samples, 62% of them with predicted HRD.

2.3 SuRe

We designed a supervised learning-based approach, called SuRe (Supervised Refitting), for signature refitting. SuRe receives as input a mutation count vector of size 96, where each entry corresponds to the number of mutation occurrences in a mutation category. Optionally, it can receive a mapping between the input samples and their corresponding tissues, expressed as a one-hot encoded vector of size T (the total number of tissues). SuRe employs a neural network model to predict signature exposures.

2.3.1 Model architecture

We propose a supervised model in which the mutation count vector of each sample i is processed to predict the relative exposures of that sample. The model's loss is the mean squared error between the predicted $\hat{y_i}$ and the actual relative exposure vector y_i of sample i (summed over all samples).

The model is based on a Mixture-of-Experts [8] neural network architecture. The architecture is depicted in Figure 1. It consists of two input layers: one input layer of size 96 (the number of categories), that handles mutation count vectors, and a second input layer of size T (the number of distinct tissues in the data set) that handles one-hot encoded tissue-type vectors. The inputs are concatenated and fed into a hidden layer with $96 + T$ neuron. This combined representation of the inputs is fed into e independent modules (experts) that are comprised of three fully-connected hidden layers, each containing h neurons. Each hidden layer applies a Rectified Linear Unit (ReLU) activation function, followed by dropout regularization, which randomly sets a fraction p of the neurons to zero during training, to prevent overfitting. The final layer of each expert module, consisting of S neurons - the number of signatures in the data set - utilizes the Softmax activation function to produce the relative exposures, a probability distribution over S signatures. The combined representation of the inputs is also fed into a gating network: a fully connected layer, containing h neurons, and a subsequent fully connected layer of size e. The output is normalized using Softmax to get the probabilities for each expert. The e module outputs are averaged with the expert probabilities as weights, to produce the final output. The rationale behind this architecture is that each expert specializes in a certain type of tissue, cancer type or exposure pattern, and the gating network decides which experts are best suited for each input.

The model was implemented in PyTorch [15], using a Stochastic Gradient Descent (SGD) optimizer and a cosine annealing with warm restarts [14] learning rate scheduler with a minimum learning rate of $\frac{1}{50}$ the initial learning rate. We implement early stopping to prevent overfitting: If the validation loss does not improve for 5 epochs, training is stopped and the weights are restored from the model that had the best performance on the validation set.

2.3.2 Training procedure and hyperparameter tuning

Our training/validation samples are randomly subsampled from a patient in the train/validation set in the following way: we randomly draw a patient i and a number m of observed mutations. m is sampled from a power-law distribution with an exponent of 1.15, resulting in a median of approximately 100 mutations when the distribution is scaled to the average number of mutations per patient in the whole genome sequencing breast cancer data set (see *Mutation data*). We then randomly draw m mutations from the sample to construct a vector of mutation counts for the sample. By following this subsampling scheme, the model is exposed during training to more sparse samples, which are inherently more difficult to predict as they often contain less information. The original WGS samples are also fed to the model during training.

We split our data sets by patients to train, validation and test sets, so that the total number of mutations in each set gives an approximate ratio of [0.7, 0.15, 0.15], respectively. For the pan-cancer data, we perform the split separately for each tissue type, so that each tissue type is represented in the three sets. The test set is not part of the training process and is used later for performance evaluation.

The parameter S is set to the number of signatures active in the input data set or the total number of signatures in the COSMIC database, and the parameter T is set to the number of tissues that are represented in the data set. We used grid search in order to tune the other two architecture parameters - e, and h. For the number of experts (e), we examined the values: $1, 4, 8, 16$, and for the other parameter, we tested $h \in \{10, 100, 500\}$.

The learning rate (r) was tuned as part of the grid search, testing the values $r \in \{0.1, 0.01, 0.001\}$. First, we tuned the three parameters on the breast cancer data set, then carried the optimal values for r to the pan-cancer data set, performing the search on e and h. For the breast cancer data set, $r = 0.1$, $h = 100$ and $e = 4$ resulted in the best performance. We note that further increasing the learning rate r and did not improve the performance on the breast cancer data set. Further tuning of e and h on the pan-cancer data set resulted in the optimal assignment of $e = 8$ and $h = 500$.

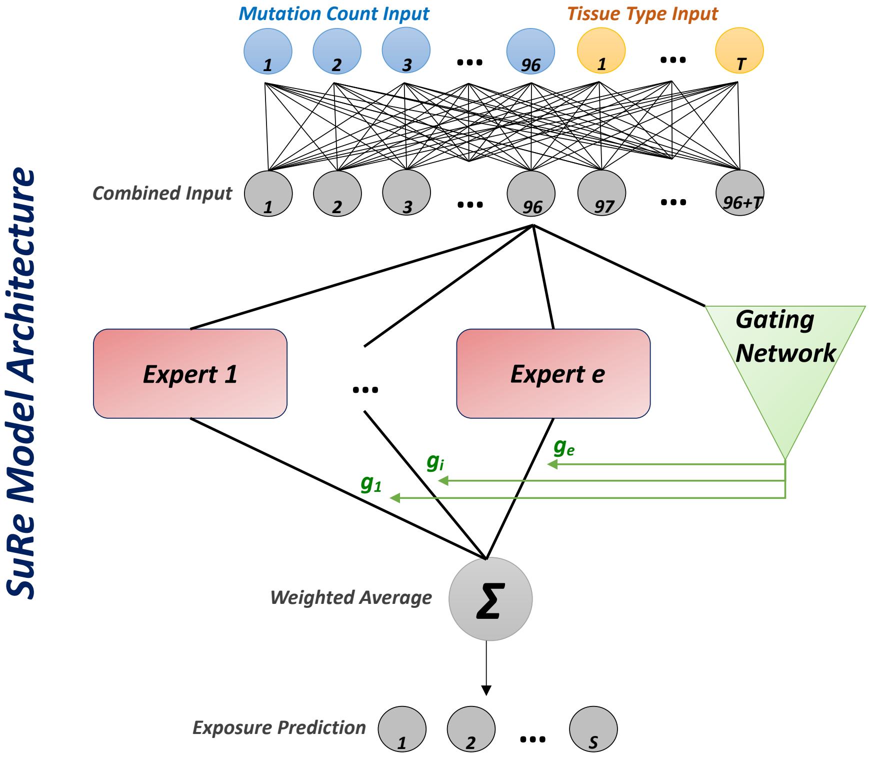


Figure 1 SuRe Mixture-of-Experts architecture. The tissue type input (size T) is concatenated to the mutation count input (96) and fed into a layer of the same size (96+T). This combined representation of the input is fed into e expert modules, as well as a gating network. Each expert outputs probabilities over S signatures, and the gating network outputs e probabilities, each corresponding to an expert. A weighted average is computed over the e expert outputs, using the gating network outputs as weights, to obtain the relative exposure prediction.

For dropout regularization, we randomly dropped 20% of the units and observed no signs of overfitting to the training data. We also tested a dropout rate of 30%, which yielded a decrease in performance.

2.4 Exposure inference assessment

We use two measures to evaluate the quality of the relative exposures predicted by each method: exposure reconstruction error and exposure correlation. Let E be the ground-truth exposures for the patients in a data set, and $\tilde{E}$ the corresponding relative exposures (i.e., normalized to sum to 1). We define the exposure reconstruction error as the L1 norm between $\tilde{E}$ and the predicted relative exposures and take the average over the samples. This metric ranges between 0 and 2: It equals 0 when the predicted relative exposures and the groud-truth exposures are identical, and equals 2 when they do not overlap at all. Similarly, we define the exposure correlation to be the average Pearson correlation over samples, i.e., correlation between the corresponding rows of $\tilde{E}$ and the predicted relative exposures. In the pan-cancer training data set, the exposure reconstruction error and the exposure correlation between two random permutations of the ground-truth exposures are 1.18 and 0.48, respectively. Lastly, we define prediction bias as the signed error, the difference between predicted and ground-truth relative exposure to a signature.

For other methods that do not directly use tissue type data, we split the samples by tissue type and inferred exposures separately for each subset of samples, based only on the signatures that are associated with that tissue, thus significantly reducing the number of signatures to be considered. For completeness, we also refitted the data to all COSMIC signatures using each competing method.

3 Results

We trained a supervised model, which we call SuRe, for signature refitting. Full details on the model and its architecture are provided in the Methods section. In the following, we conduct comprehensive testing of our method on synthetic and real data sets and demonstrate its clinical applications. SuRe is trained using COSMIC exposures for whole genome sequencing samples (real and synthetic), as targets. The COSMIC exposures were derived by refitting WGS mutation data to the database of known signatures using SigProfiler.

In order to test our algorithm's performance under varying numbers of mutations per sample, we further constructed simulated data sets in which we downsampled the original data by randomly sampling $m \in \{300, 100, 30, 10, 6, 3\}$ mutations per patient without replacement.

We compare SuRe's performance against two leading refitting methods: SigProfiler-Assignment and MuSiCal, as well as Mix - a method specialized in the refitting of sparse mutation data, and the popular non-negative least squares (NNLS) method that minimizes the Kullback-Leibler divergence. We could not include SigNet in the comparison as it failed to install due to an unavailable package requirement.

3.1 Performance evaluation on synthetic data

As a first test of our approach, we evaluated SuRe using synthetically generated samples that were previously used for benchmarking of both de novo extraction [11] and refitting [7] scenarios. This data set allows us to compare all methods against ground-truth exposures. Interestingly, the top-performing method on the full set of genome-wide mutations is MuSiCal. SuRe outperforms all methods on both metrics when there are less than 300 mutations per sample. As expected, all methods performed worse when given fewer input mutations, although SuRe and Mix were more robust to sparse data compared to NNLS, MuSiCal and SigProfiler. These results are summarized in Figure 2. Supplementary Figure 6 shows the results when the same synthetic data set is refitted to all COSMIC signatures, instead of only the signatures that are known to be active in the data set. We see that NNLS, MuSiCal and SigProfiler perform much worse in this case, as they suffer from over-assignment of exposures to signatures that are inactive.

3.2 Prediction of COSMIC exposures in breast cancer

As a second test of our approach, we applied it to analyze real samples originating from a single cancer type, focusing on breast cancer which had the highest number of samples. We further restricted attention to samples that underwent whole-genome-sequencing as we reasoned that the COSMIC estimated exposures for these samples were the most accurate. Typically, whole-genome-sequencing (WGS) yields high mutation counts (thousands of mutations), leading to larger mutation sample sizes, lower sampling variation, and greater confidence in

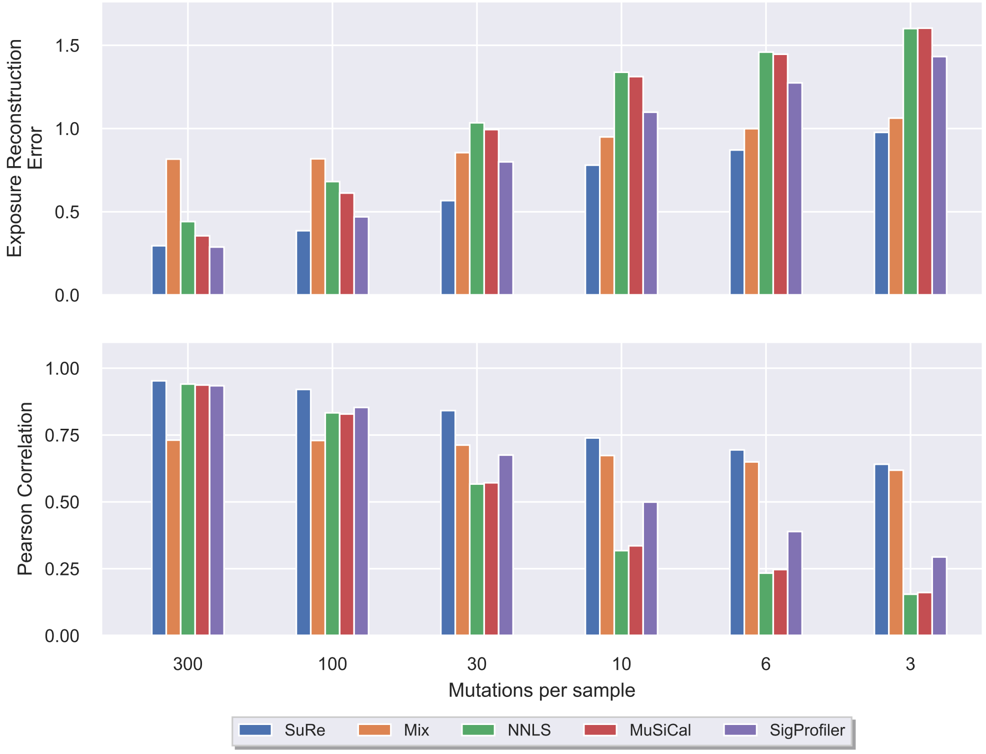


Figure 2 Performance evaluation on synthetic data. Each bar represents the mean value over $\lceil \frac{300}{m} \rceil$ repetitions, where m is the number mutations per sample. The same m mutations are used to evaluate all methods in each repetition.

inferring the underlying mutation distribution. In this analysis, we used only mutational signatures that were active in at least one of the patients in the data set according to COSMIC. We also provide results for the scenario where the data is refitted to all signatures, using all competing methods.

The results are summarized in Figure 3 and show that the exposures predicted by SuRe were in better agreement with COSMIC, in terms of reconstruction error and correlation across all sampling sizes. The results when the breast cancer data is refitted to all COSMIC signatures are provided in Supplementary Figure 7, and again demonstrate that SuRe is more resilient to the challenges of refitting to a large database of potentially inactive signatures.

More recently, a re-analysis of COSMIC samples was published based on a new and improved version of the SigProfiler software [11]. This re-analysis runs sigProfileExtractor for de novo extraction of mutational signatures, then refits the mutation counts to these signatures using SigProfilerAssignment in order to infer sample exposures. We re-trained SuRe using these predicted exposures. In this case, the exposures predicted by *SigProfiler* on the full WGS mutation counts represent the " ground-truth" exposures, but the algorithm's predictions on downsampled instances could deviate from them. The results summarized in Supplementary Figure 8 show that exposures predicted by SuRe are in better agreement with the new COSMIC exposures compared to other methods (excluding SigProfiler) on all sampling sizes. Interestingly, SuRe is comparable in performance to SigProfiler for 300

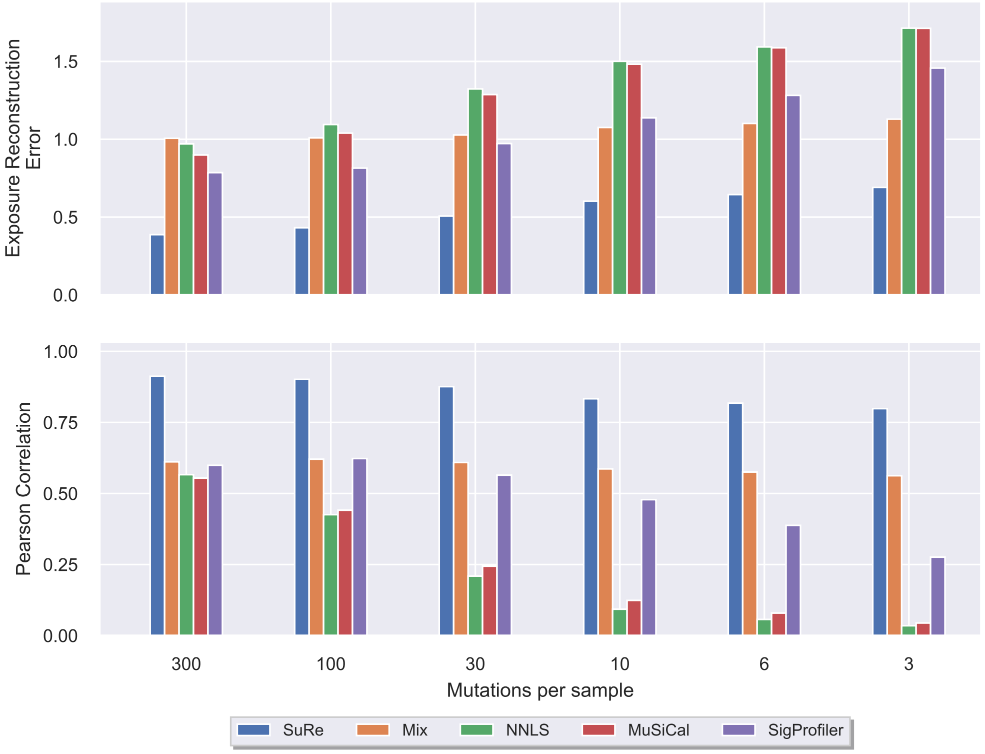


Figure 3 Comparative assessment of exposure prediction in breast cancer patients. Each value represents the mean over $\lceil \frac{300}{m} \rceil$ repetitions, where m is the number mutations per sample. The same m mutations are used to evaluate all methods in each repetition.

mutations, and outperforms it when there are fewer mutations per sample. As in previous results, the gap in performance increases when the refitting is done with respect to all signatures (Supplementary Figure 9).

3.3 A pan-cancer application

Next, we wanted to evaluate SuRe in a pan-cancer setting. Typically, exposure refitting is performed using a limited set of signatures that are believed to be active in the (type-specific) data set, based on prior knowledge. This is the case since the inference is more challenging as the number of mutational signatures to be considered in the refitting is greater (in particular, there are 58 active signatures in the pan-cancer data set vs. 20 in breast cancer data set), and often results in over-assignment of non-zero exposures to inactive signatures, as illustrated in [12]. However, reliably taking into account a broader set of signatures could potentially reveal the presence of signatures in a patient, which would otherwise have been overlooked.

For pan-cancer refitting, SuRe is capable of leveraging the samples' cancerous tissue type (one of 14 ICGC tissue types that are available in the data set), which it receives as input, to guide the assignment toward signatures that are more likely to be active, while attempting to refit the mutation data to all 58 signatures. The results on the pan-cancer data set are summarized in Figure 4, and show a clear advantage for SuRe compared to previous methods. As expected, previous methods achieve better performance when refitting is based only on signatures that are known to be active in the corresponding tissue type, rather than based on all 58 signatures. The distributions of reconstruction errors and correlations for each method across tissue types are summarized in Supplementary Figure 10. To assess these results in a more specific way, we focus on a midrange downsampling size of 30 mutations for the following analyses (Supplementary Figures 11 and 12). SuRe is the highest scoring method on all tissue types, when previous methods are refitted to all signatures. When each of the previous methods is applied separately for each tissue type, SuRe performs best on the vast majority of tissue types, despite being the only method not explicitly refitted to signatures that are known to be active in the tissue. Next, we wanted to examine whether the different characteristics of the cancer-specific data sets correlated with the difference in performance between SuRe and the other method that performed best on each cancer type. Hence, we took the difference between the reconstruction error of SuRe and of the competing method and computed its correlation to several features. We found that the number of samples and the total number of active signatures in a data set were positively correlated with the improvement of SuRe over the other top performer (0.41 and 0.53, respectively). This was also true for the fraction of flat signatures in the data (0.56). To a lesser extent, the fraction of rare signatures (occur in less than 10% samples) also showed a small positive correlation (0.17). Finally, to evaluate the importance of tissue-type input data in the performance of SuRe, we also trained it without providing it with this data. Although this version still outperforms previous methods in terms of reconstruction error and correlation to COSMIC exposures, we observe a clear drop in the model's capability when tissue-type data are omitted. These results are summarized in Supplementary Figure 16.

3.4 Prediction of clinical attributes

To test the utility of SuRe in a clinical scenario, we examined its ability to predict homologous recombination deficiency (HRD) status from panel sequencing data. Starting from triple negative breast cancer samples with known HRD status [19], we downsampled the mutation data to the genomic regions of the MSK-IMPACT panel [4]. We then applied SuRe and competing methods to analyze the downsampled data. Since COSMIC's SBS3 signature is known to correlate extremely well with HRD status [19], we used the inferred SBS3 exposure to estimate HRD. That is, the predicted SBS3 exposures according to each method were used to classify the samples as HRD positive or HRD negative based on different decision thresholds. In order to fully leverage our model capabilities, we fine-tuned SuRe to specifically predict SBS3 exposure by adjusting its loss to be the mean squared error between the predicted exposure to SBS3 and the actual exposure to SBS3. Figure 5 shows the ROC curves of the prediction using each method. As a baseline, we added a classifier that is based solely on the total mutation burden in each sample. SuRe and Mix were the only methods that clearly improved over this baseline, and SuRe significantly outperformed Mix.

To further examine these results, we tested the SBS3 prediction bias of each method on the breast cancer data set. To that end, we computed for each sample the difference between the predicted relative exposure to SBS3 and the relative exposure according to COSMIC. We found that other methods tend to underestimate exposure to SBS3 (Supplementary Figure 13). This phenomenon was also apparent in SBS5, the other flat signature common in breast cancer samples, but not observed in spiky signatures that are common in breast cancer (SBS1, SBS2, SBS13, SBS18, SBS34). In comparison, SuRe does not show a tendency to underestimate or overestimate these signatures (Supplementary Figures 14 and 15).

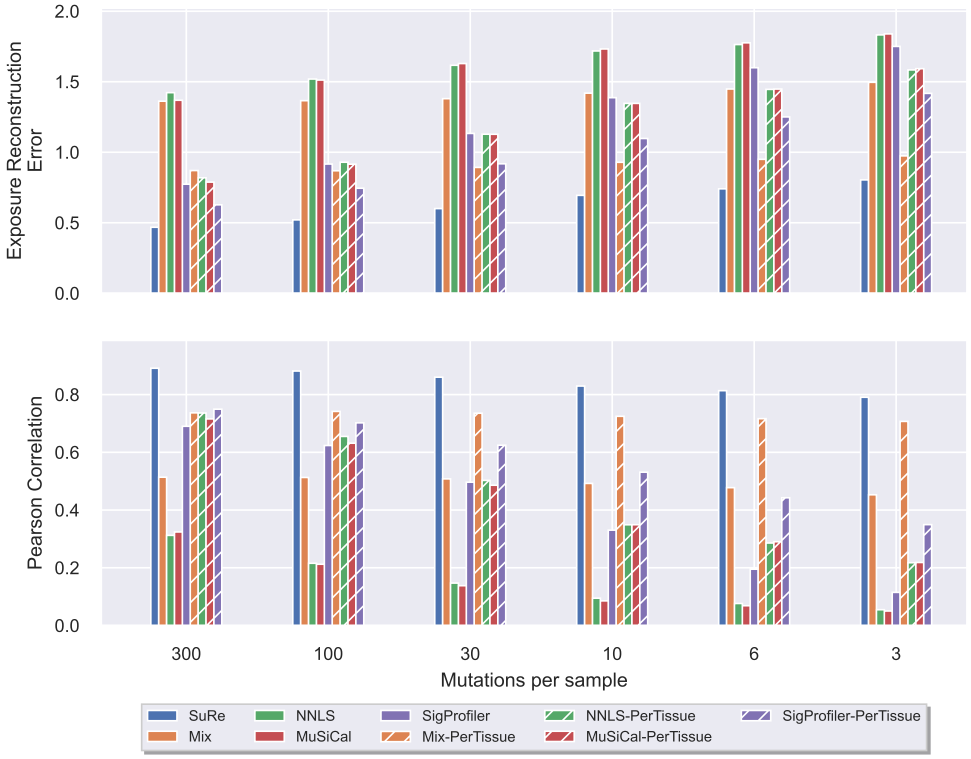


Figure 4 Comparative assessment - targeting COSMIC exposures on pan-cancer patients. Each bar represents the mean value over $\lceil\frac{300}{m}\rceil$ repetitions, where m is the number mutations per sample. The same m mutations are used to evaluate all methods in each repetition.

4 Conclusions

We have presented SuRe, a supervised method for mutational signature refitting on sparse mutation data.

One potential limitation of SuRe is the use of COSMIC exposures as labels for its training and evaluation. These exposures were estimated using the SigProfilerAttribution algorithm [2, 11]. Clearly, the use of computationally predicted exposures as ground-truth labels could restrict and bias the learning process. Nevertheless, we opted to train SuRe based on the COSMIC exposures, as they serve as a standard in the field and were shown to correlate with different clinical attributes, including patient age [3, 2], tobacco smoking history [1], homologous recombination deficiency status [5] and mismatch repair deficiency [3]. Reassuringly, the ability of SuRe to better reconstruct exposures that had been estimated using rich data from very few mutations suggests that it can successfully leverage information of signature co-activity, which compensates for the sparse input data. This notion is reaffirmed by SuRe's success compared to other methods in predicting homologous recombination deficiency from sparse panel sequencing data. These results suggest that SuRe could be leveraged to more accurately analyze large cohorts of targeted sequencing data.

SuRe can be easily fine-tuned to predict exposures of specific signatures, or subsets of signatures that correspond to specific cancer types or biological mechanisms, by adjusting its loss function so that the error for these signatures is minimized. In addition, SuRe provides

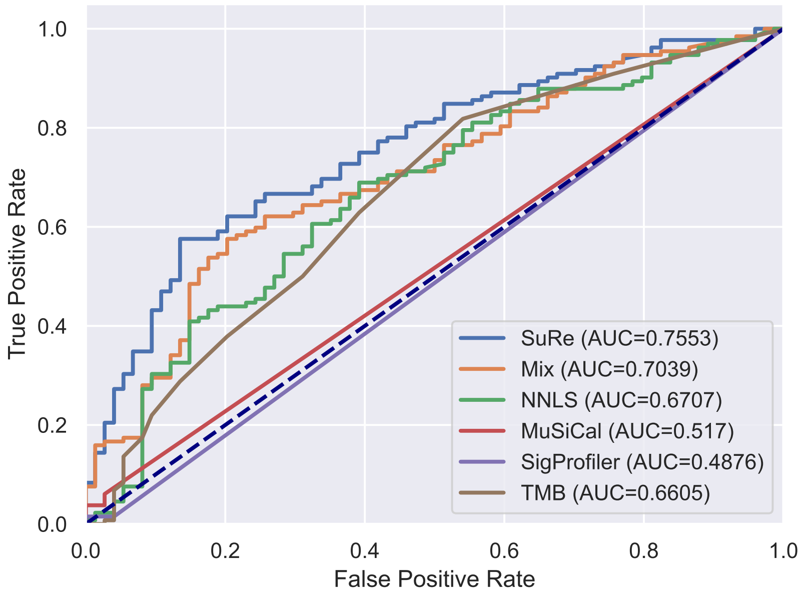


Figure 5 ROC curves for HR deficiency prediction - SBS3-optimized breast cancer model.

a major improvement over previous methods in the handling of pan-cancer data, which could open the door for reliable refitting to broader sets of signatures, thereby revealing the activity of rare signatures in a patient.

References

1 Ludmil B Alexandrov, Young Seok Ju, Kerstin Haase, Peter Van Loo, Iñigo Martincorena, Serena Nik-Zainal, Yasushi Totoki, Akihiro Fujimoto, Hidewaki Nakagawa, Tatsuhiro Shibata, Peter J Campbell, Paolo Vineis, David H Phillips, and Michael R Stratton. Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312):618–622, November 2016.

2 Ludmil B. Alexandrov, Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, Kyle R. Covington, Dmitry A. Gordenin, Erik N. Bergstrom, S. M. Ashiqul Islam, Nuria Lopez-Bigas, Leszek J. Klimczak, John R. McPherson, Sandro Morganella, Radhakrishnan Sabarinathan, David A. Wheeler, Ville Mustonen, Paul Boutros, Kin Chan, Akihiro Fujimoto, Gad Getz, Marat Kazanov, Michael Lawrence, Iñigo Martincorena, Hidewaki Nakagawa, Paz Polak, Stephenie Prokopec, Steven A. Roberts, Steven G. Rozen, Natalie Saini, Tatsuhiro Shibata, Yuichi Shiraishi, Michael R. Stratton, Bin Tean Teh, Ignacio Vázquez-García, Fouad Yousif, Willie Yu, Lauri A. Aaltonen, Federico Abascal, Adam Abeshouse, Hiroyuki Aburatani, David J. Adams, Nishant Agrawal, Keun Soo Ahn, Sung-Min Ahn, Hiroshi Aikata, Rehan Akbani, Kadir C. Akdemir, Hikmat Al-Ahmadie, Sultan T. Al-Sedairy, Fatima Al-Shahrour, Malik Alawi, Monique Albert, Kenneth Aldape, Adrian Ally, Kathryn Alsop, Eva G. Alvarez, Fernanda Amary, Samirkumar B. Amin, Brice Aminou, Ole Ammerpohl, Matthew J. Anderson, Yeng Ang, Davide Antonello, Pavana Anur, Samuel Aparicio, Elizabeth L. Appelbaum, Yasuhito Arai, Axel Aretz, Koji Arihiro, Shun-ichi Ariizumi, Joshua Armenia, Laurent Arnould, Sylvia Asa, Yassen Assenov, Gurnit Atwal, Sietse Aukema, J. Todd Auman, Miriam R. R. Aure, Philip Awadalla, Marta Aymerich, Gary D. Bader, Adrian Baez-Ortega, Matthew H. Bailey, Peter J. Bailey, Miruna Balasundaram, Saianand Balu, Pratiti Bandopadhayay, Rosamonde E. Banks, Stefano Barbi, Andrew P. Barbour, Jonathan Barenboim, Jill Barnholtz-Sloan, Hugh Barr, Elisabet Barrera, John

Bartlett, Javier Bartolome, Claudio Bassi, Oliver F. Bathe, Daniel Baumhoer, Prashant Bavi, Stephen B. Baylin, Wojciech Bazant, Duncan Beardsmore, Timothy A. Beck, Sam Behjati, Andreas Behren, Beifang Niu, Cindy Bell, Sergi Beltran, Christopher Benz, Andrew Berchuck, Anke K. Bergmann, Benjamin P. Berman, Daniel M. Berney, Stephan H. Bernhart, Rameen Beroukhim, Mario Berrios, Samantha Bersani, Johanna Bertl, Miguel Betancourt, Vinayak Bhandari, Shriram G. Bhosle, Andrew V. Biankin, Matthias Bieg, Darell Bigner, Hans Binder, Ewan Birney, Michael Birrer, Nidhan K. Biswas, Bodil Bjerkehagen, Tom Bodenheimer, Lori Boice, Giada Bonizzato, Johann S. De Bono, Moiz S. Bootwalla, Ake Borg, Arndt Borkhardt, Keith A. Boroevich, Ivan Borozan, Christoph Borst, Marcus Bosenberg, Mattia Bosio, Jacqueline Boultwood, Guillaume Bourque, Paul C. Boutros, G. Steven Bova, David T. Bowen, Reanne Bowlby, David D. L. Bowtell, Sandrine Boyault, Rich Boyce, Jeffrey Boyd, Alvis Brazma, Paul Brennan, Daniel S. Brewer, Arie B. Brinkman, Robert G. Bristow, Russell R. Broaddus, Jane E. Brock, Malcolm Brock, Annegien Broeks, Angela N. Brooks, Denise Brooks, Benedikt Brors, Søren Brunak, Timothy J. C. Bruxner, Alicia L. Bruzos, Alex Buchanan, Ivo Buchhalter, Christiane Buchholz, Susan Bullman, Hazel Burke, Birgit Burkhardt, Kathleen H. Burns, John Busanovich, Carlos D. Bustamante, Adam P. Butler, Atul J. Butte, Niall J. Byrne, Anne-Lise Børresen-Dale, Samantha J. Caesar-Johnson, Andy Cafferkey, Declan Cahill, Claudia Calabrese, Carlos Caldas, Fabien Calvo, Niedzica Camacho, Peter J. Campbell, Elias Campo, Cinzia Cantù, Shaolong Cao, Thomas E. Carey, Joana Carlevaro-Fita, Rebecca Carlsen, Ivana Cataldo, Mario Cazzola, Jonathan Cebon, Robert Cerfolio, Dianne E. Chadwick, Dimple Chakravarty, Don Chalmers, Calvin Wing Yiu Chan, Michelle Chan-Seng-Yue, Vishal S. Chandan, David K. Chang, Stephen J. Chanock, Lorraine A. Chantrill, Aurélien Chateigner, Nilanjan Chatterjee, Kazuaki Chayama, Hsiao-Wei Chen, Jieming Chen, Ken Chen, Yiwen Chen, Zhaohong Chen, Andrew D. Cherniack, Jeremy Chien, Yoke-Eng Chiew, Suet-Feung Chin, Juok Cho, Sunghoon Cho, Jung Kyoon Choi, Wan Choi, Christine Chomienne, Zechen Chong, Su Pin Choo, Angela Chou, Angelika N. Christ, Elizabeth L. Christie, Eric Chuah, Carrie Cibulskis, Kristian Cibulskis, Sara Cingarlini, Peter Clapham, Alexander Claviez, Sean Cleary, Nicole Cloonan, Marek Cmero, Colin C. Collins, Ashton A. Connor, Susanna L. Cooke, Colin S. Cooper, Leslie Cope, Vincenzo Corbo, Matthew G. Cordes, Stephen M. Cordner, Isidro Cortés-Ciriano, Kyle Covington, Prue A. Cowin, Brian Craft, David Craft, Chad J. Creighton, Yupeng Cun, Erin Curley, Ioana Cutcutache, Karolina Czajka, Bogdan Czerniak, Rebecca A. Dagg, Ludmila Danilova, Maria Vittoria Davi, Natalie R. Davidson, Helen Davies, Ian J. Davis, Brandi N. Davis-Dusenbery, Kevin J. Dawson, Francisco M. De La Vega, Ricardo De Paoli-Iseppi, Timothy Defreitas, Angelo P. Dei Tos, Olivier Delaneau, John A. Demchok, PCAWG Mutational Signatures Working Group, and P. C. A. W. G. Consortium. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, February 2020. `doi:10.1038/s41586-020-1943-3`.

3 Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge, Peter J. Campbell, and Michael R. Stratton. Deciphering signatures of mutational processes operative in human cancer. *Cell Reports*, 3(1):246–259, January 2013. `doi:10.1016/j.celrep.2012.12.008`.

4 Donavan T. Cheng, Talia N. Mitchell, Ahmet Zehir, Ronak H. Shah, Ryma Benayed, Aijazuddin Syed, Raghu Chandramohan, Zhen Yu Liu, Helen H. Won, Sasinya N. Scott, A. Rose Brannon, Catherine O'Reilly, Justyna Sadowska, Jacklyn Casanova, Angela Yannes, Jaclyn F. Hechtman, Jinjuan Yao, Wei Song, Dara S. Ross, Alifya Oultache, Snjezana Dogan, Laetitia Borsu, Meera Hameed, Khedoudja Nafa, Maria E. Arcila, Marc Ladanyi, and Michael F. Berger. Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (msk-impact): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *The Journal of Molecular Diagnostics*, 17(3):251–264, May 2015. `doi:10.1016/j.jmoldx.2014.12.006`.

5 Helen Davies, Dominik Glodzik, Sandro Morganella, Lucy R. Yates, Johan Staaf, Xueqing Zou, Manasa Ramakrishna, Sancha Martin, Sandrine Boyault, Anieta M. Sieuwerts, Peter T. Simpson, Tari A. King, Keiran Raine, Jorunn E. Eyfjord, Gu Kong, Åke Borg, Ewan Birney,

Hendrik G. Stunnenberg, Marc J. van de Vijver, Anne-Lise Børresen-Dale, John W. M. Martens, Paul N. Span, Sunil R. Lakhani, Anne Vincent-Salomon, Christos Sotiriou, Andrew Tutt, Alastair M. Thompson, Steven Van Laere, Andrea L. Richardson, Alain Viari, Peter J. Campbell, Michael R. Stratton, and Serena Nik-Zainal. Hrdetect is a predictor of brca1 and brca2 deficiency based on mutational signatures. *Nature Medicine*, 23(4):517–525, April 2017. doi:10.1038/nm.4292.

6 Andrea Degasperi, Xueqing Zou, Tauanne Dias Amarante, Andrea Martinez-Martinez, Gene Ching Chiek Koh, João M. L. Dias, Laura Heskin, Lucia Chmelova, Giuseppe Rinaldi, Valerie Ya Wen Wang, Arjun S. Nanda, Aaron Bernstein, Sophie E. Momen, Jamie Young, Daniel Perez-Gil, Yasin Memari, Cherif Badja, Scott Shooter, Jan Czarnecki, Matthew A. Brown, Helen R. Davies, null null, Serena Nik-Zainal, J. C. Ambrose, P. Arumugam, R. Bevers, M. Bleda, F. Boardman-Pretty, C. R. Boustred, H. Brittain, M. J. Caulfield, G. C. Chan, T. Fowler, A. Giess, A. Hamblin, S. Henderson, T. J. P. Hubbard, R. Jackson, L. J. Jones, D. Kasperaviciute, M. Kayikci, A. Kousathanas, L. Lahnstein, S. E. A. Leigh, I. U. S. Leong, F. J. Lopez, F. Maleady-Crowe, M. McEntagart, F. Minneci, L. Moutsianas, M. Mueller, N. Murugaesu, A. C. Need, P. O'Donovan, C. A. Odhams, C. Patch, D. Perez-Gil, M. B. Pereira, J. Pullinger, T. Rahim, A. Rendon, T. Rogers, K. Savage, K. Sawant, R. H. Scott, A. Siddiq, A. Sieghart, S. C. Smith, A. Sosinsky, A. Stuckey, M. Tanguy, A. L. Taylor Tavares, E. R. A. Thomas, S. R. Thompson, A. Tucci, M. J. Welland, E. Williams, K. Witkowska, and S. M. Wood. Substitution mutational signatures in whole-genome-sequenced cancers in the uk population. *Science*, 376(6591):abl9283, 2022. doi:10.1126/science.abl9283.

7 Marcos Díaz-Gay, Raviteja Vangara, Mark Barnes, Xi Wang, S M Ashiqul Islam, Ian Vermes, Nithish Bharadhwaj Narasimman, Ting Yang, Zichen Jiang, Sarah Moody, Sergey Senkin, Paul Brennan, Michael R Stratton, and Ludmil B Alexandrov. Assigning mutational signatures to individual samples and individual somatic mutations with sigprofilerassignment. *bioRxiv*, 2023. doi:10.1101/2023.07.10.548264.

8 David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts, 2014. arXiv:1312.4314.

9 W. Morven Gentleman. Solving least squares problems (charles l. lawson and richard j. hanson). *SIAM Review*, 18(3):518–520, 1976. doi:10.1137/1018100.

10 Xiaoqing Huang, Damian Wojtowicz, and Teresa M Przytycka. Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics*, 34(2):330–337, September 2017. doi:10.1093/bioinformatics/btx604.

11 S.M. Ashiqul Islam, Marcos Díaz-Gay, Yang Wu, Mark Barnes, Raviteja Vangara, Erik N. Bergstrom, Yudou He, Mike Vella, Jingwei Wang, Jon W. Teague, Peter Clapham, Sarah Moody, Sergey Senkin, Yun Rose Li, Laura Riva, Tongwu Zhang, Andreas J. Gruber, Christopher D. Steele, Burçak Otlu, Azhar Khandekar, Ammal Abbasi, Laura Humphreys, Natalia Syulyukina, Samuel W. Brady, Boian S. Alexandrov, Nischalan Pillay, Jinghui Zhang, David J. Adams, Iñigo Martincorena, David C. Wedge, Maria Teresa Landi, Paul Brennan, Michael R. Stratton, Steven G. Rozen, and Ludmil B. Alexandrov. Uncovering novel mutational signatures by de novo extraction with sigprofilerextractor. *Cell Genomics*, 2(11):100179, 2022. doi:10.1016/j.xgen.2022.100179.

12 Hu Jin, Doga C. Gulhan, Daniel Ben-Isvy, David Geng, Viktor Ljungstrom, and Peter J. Park. Accurate and sensitive mutational signature analysis with musical. *bioRxiv*, 2022. doi:10.1101/2022.04.21.489082.

13 Shantao Li, Forrest W. Crawford, and Mark B. Gerstein. Using siglasso to optimize cancer mutation signatures jointly with sampling likelihood. *Nature Communications*, 11(1):3575, July 2020. doi:10.1038/s41467-020-17388-x.

14 Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. arXiv:1608.03983.

15 Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

16 Rachel Rosenthal, Nicholas McGranahan, Javier Herrero, Barry S. Taylor, and Charles Swanton. deconstructsigs: delineating mutational processes in single tumors distinguishes dna repair deficiencies and patterns of carcinoma evolution. *Genome Biology*, 17(1):31, February 2016. doi:10.1186/s13059-016-0893-4.

17 Itay Sason, Yuexi Chen, Mark D.M. Leiserson, and Roded Sharan. A mixture model for signature discovery from sparse mutation data. *Genome Medicine*, 13(1):173, November 2021. doi:10.1186/s13073-021-00988-7.

18 Claudia Serrano Colome, Oleguer Canal Anton, Vladimir Seplyarskiy, and Donate Weghorn. Mutational signature decomposition with deep neural networks reveals origins of clock-like processes and hypoxia dependencies. *bioRxiv*, 2023. doi:10.1101/2023.12.06.570467.

19 Johan Staaf, Dominik Glodzik, Ana Bosch, Johan Vallon-Christersson, Christel Reuterswärd, Jari Häkkinen, Andrea Degasperi, Tauanne Dias Amarante, Lao H. Saal, Cecilia Hegardt, Hilary Stobart, Anna Ehinger, Christer Larsson, Lisa Rydén, Niklas Loman, Martin Malmberg, Anders Kvist, Hans Ehrencrona, Helen R. Davies, Åke Borg, and Serena Nik-Zainal. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nature Medicine*, 25(10):1526–1533, October 2019. doi:10.1038/s41591-019-0582-4.

20 John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C Ramshaw, Claire E Rye, Helen E Speedy, Ray Stefancsik, Sam L Thompson, Shicai Wang, Sari Ward, Peter J Campbell, and Simon A Forbes. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1):D941–D947, October 2018. doi:10.1093/nar/gky1015.

A Supplementary Material

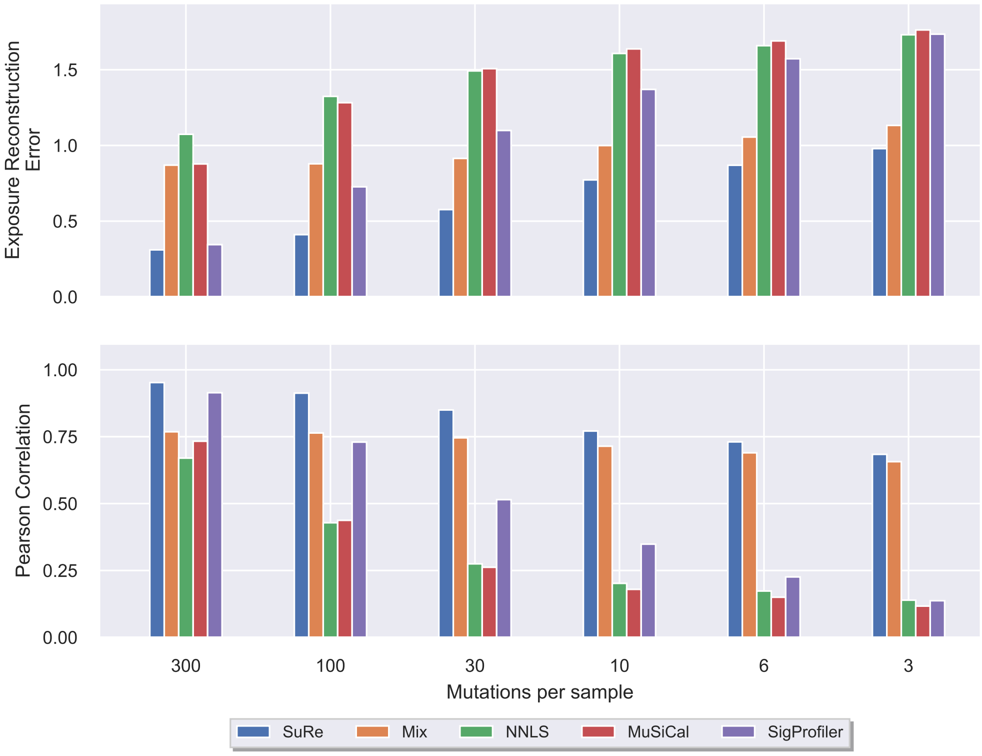


Figure 6 Performance evaluation on synthetic data - refitting to all signatures. Each bar represents the mean value over $\lceil \frac{300}{m} \rceil$ repetitions, where m is the number mutations per sample. The same m mutations are used to evaluate all methods in each repetition.

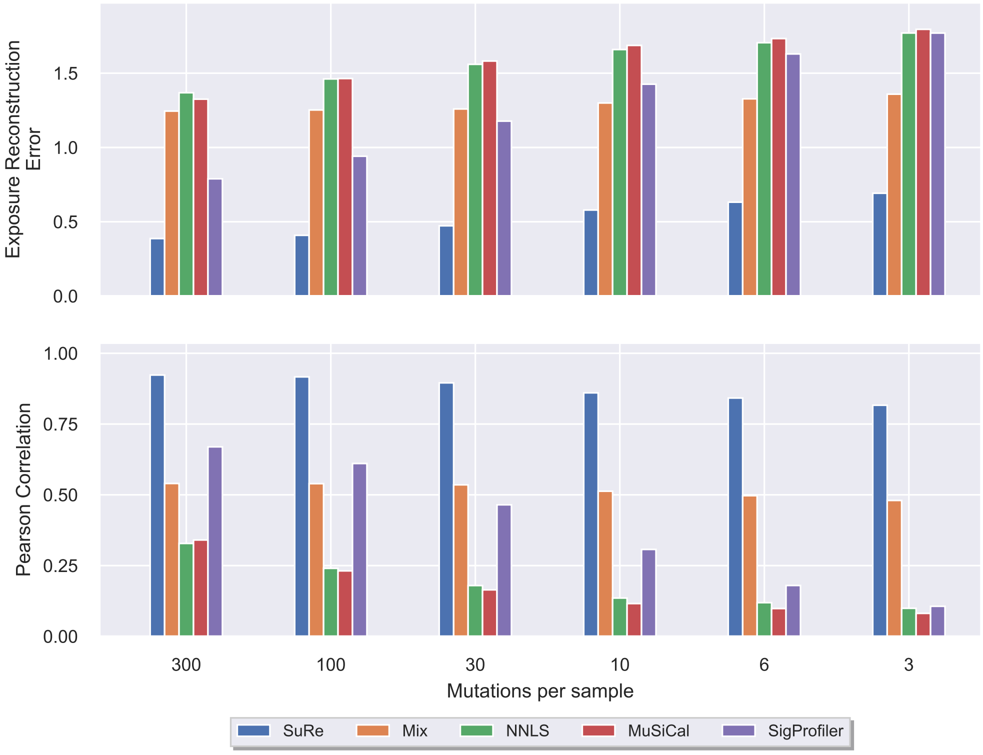


Figure 7 Comparative assessment of exposure prediction in breast cancer data - refitting to all signatures. Each value represents the mean over $\lceil \frac{300}{m} \rceil$ repetitions, where m is the number mutations per sample. The same m mutations are used to evaluate all methods in each repetition.

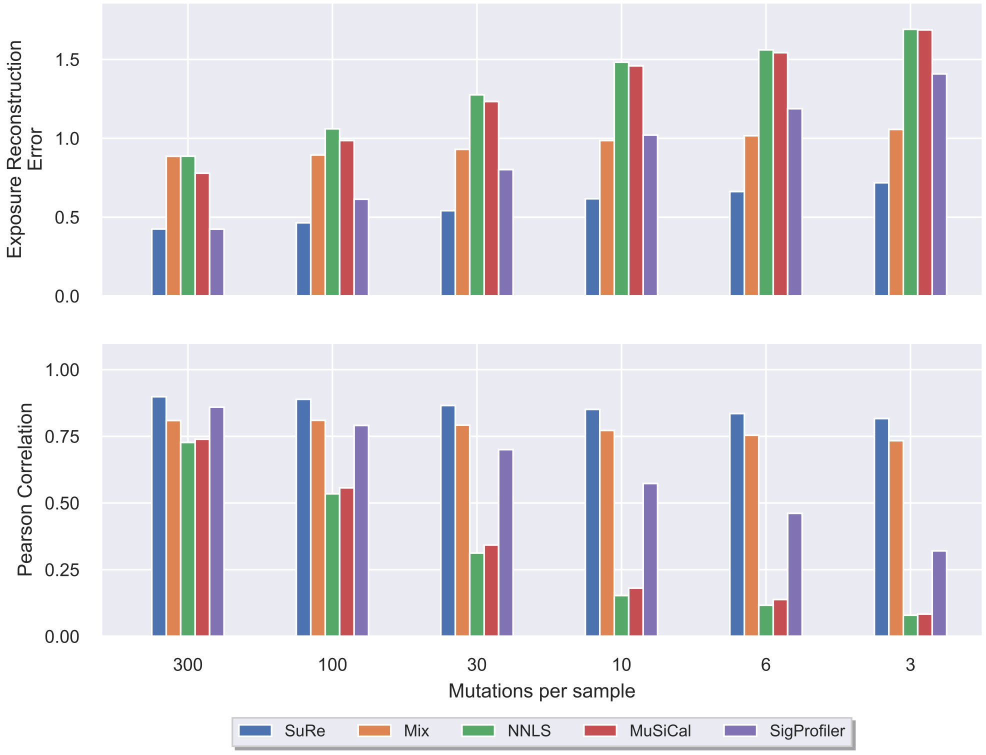


Figure 8 Comparative assessment of exposure prediction in reanalyzed breast cancer data - refitting to active signatures. Each value represents the mean over $\lceil \frac{300}{m} \rceil$ repetitions, where m is the number mutations per sample. The same m mutations are used to evaluate all methods in each repetition.

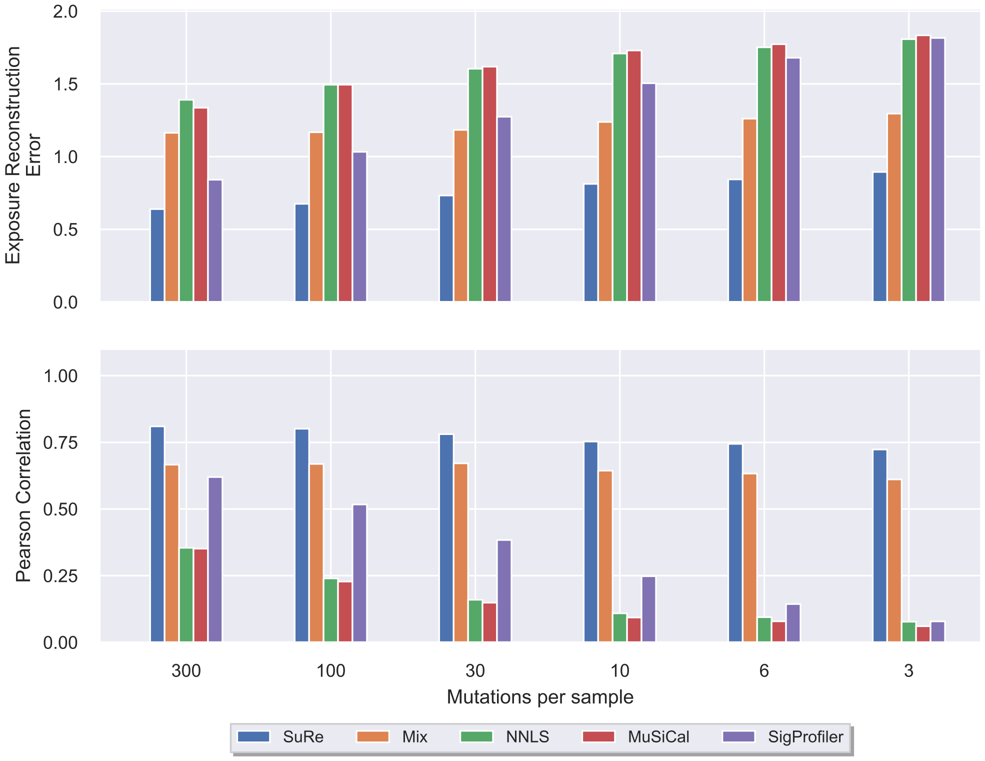


Figure 9 Comparative assessment of exposure prediction in reanalyzed breast cancer data - refitting to all signatures. Each value represents the mean over $\lceil \frac{300}{m} \rceil$ repetitions, where m is the number mutations per sample. The same m mutations are used to evaluate all methods in each repetition.

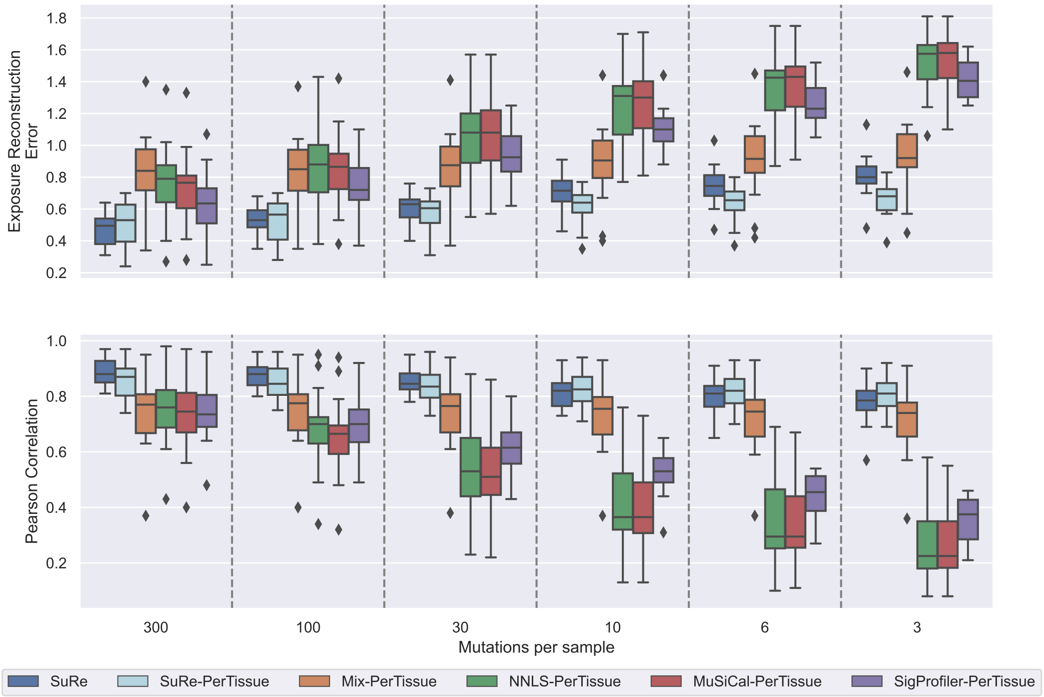
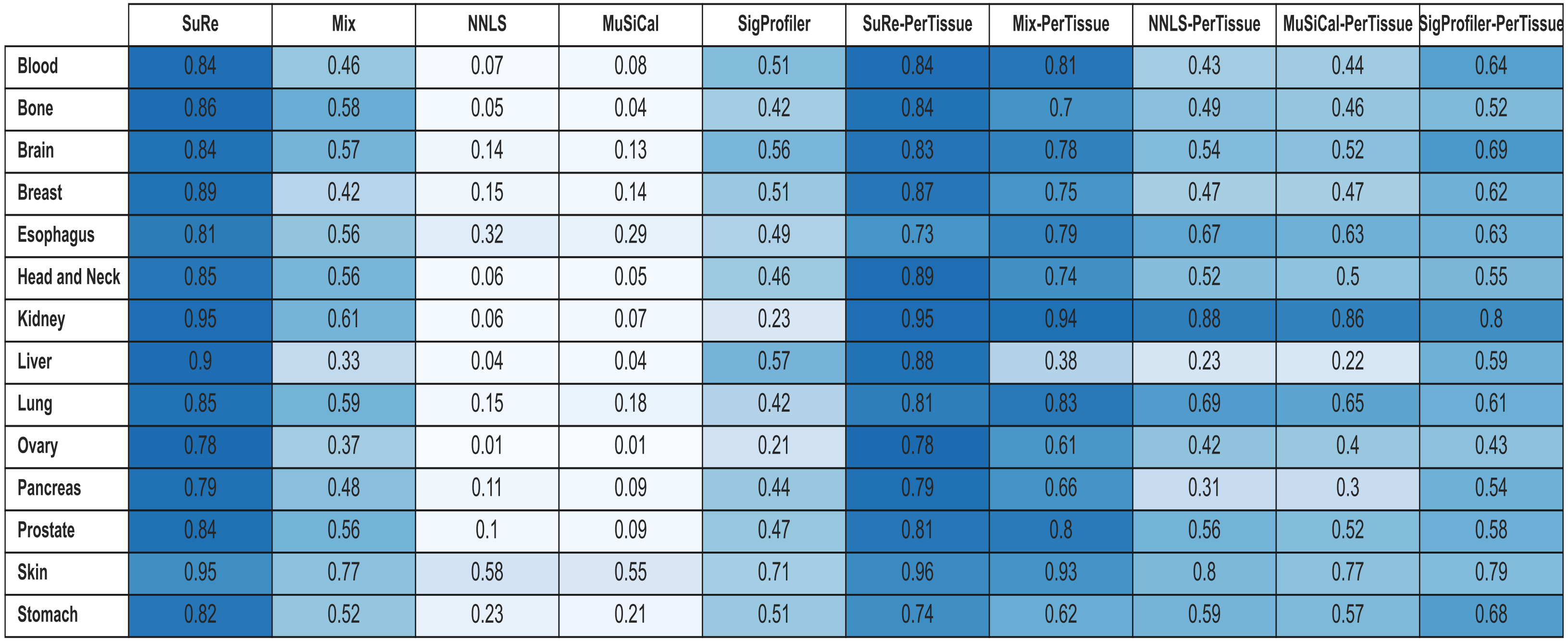


Figure 10 Comparative assessment on pan-cancer patients - distribution across tissue types. Each box represents the values across the 14 tissue types. The black line within each whisker represents to the median value. Whiskers span from the 25th to the 75th percentile.

	SuRe	Mix	NNLS	MuSiCal	SigProfiler	SuRe-PerTissue	Mix-PerTissue	NNLS-PerTissue	MuSiCal-PerTissue	SigProfiler-PerTissue
Blood	0.65	1.5	1.73	1.72	1.14	0.59	1.0	1.32	1.28	0.94
Bone	0.57	1.39	1.73	1.74	1.21	0.55	0.91	1.12	1.13	1.07
Brain	0.62	1.45	1.63	1.65	1.02	0.59	0.86	1.09	1.07	0.73
Breast	0.53	1.38	1.56	1.57	1.1	0.5	0.89	1.14	1.12	0.92
Esophagus	0.64	1.19	1.35	1.38	1.05	0.65	0.63	0.79	0.84	0.88
Head and Neck	0.54	1.36	1.7	1.71	1.16	0.44	0.83	1.07	1.09	1.06
Kidney	0.4	1.41	1.81	1.81	1.5	0.31	0.37	0.55	0.57	0.62
Liver	0.62	1.63	1.81	1.81	1.16	0.64	1.41	1.57	1.57	1.07
Lung	0.66	1.34	1.71	1.67	1.25	0.63	0.72	0.87	0.89	0.93
Ovary	0.75	1.48	1.8	1.8	1.5	0.68	1.03	1.22	1.25	1.25
Pancreas	0.76	1.43	1.65	1.66	1.2	0.73	1.07	1.38	1.37	1.05
Prostate	0.66	1.42	1.69	1.7	1.13	0.62	0.81	1.04	1.07	0.92
Skin	0.4	0.72	1.07	1.12	0.86	0.36	0.41	0.75	0.78	0.72
Stomach	0.67	1.35	1.48	1.5	1.07	0.71	0.97	0.95	0.95	0.82

Figure 11 Comparative assessment on pan-cancer patients - reconstruction error per tissue type breakdown. Each entry represents the mean reconstruction error value between the predicted exposures and the COSMIC exposures over 10 repetitions of the experiment. In each experiment, 30 mutations are randomly sampled per patient and the relative signature exposures are predicted using each method. The same 30 mutations are used to evaluate all methods in each repetition.



	SuRe	Mix	NNLS	MuSiCal	SigProfiler	SuRe-PerTissue	Mix-PerTissue	NNLS-PerTissue	MuSiCal-PerTissue	SigProfiler-PerTissue
Blood	0.84	0.46	0.07	0.08	0.51	0.84	0.81	0.43	0.44	0.64
Bone	0.86	0.58	0.05	0.04	0.42	0.84	0.7	0.49	0.46	0.52
Brain	0.84	0.57	0.14	0.13	0.56	0.83	0.78	0.54	0.52	0.69
Breast	0.89	0.42	0.15	0.14	0.51	0.87	0.75	0.47	0.47	0.62
Esophagus	0.81	0.56	0.32	0.29	0.49	0.73	0.79	0.67	0.63	0.63
Head and Neck	0.85	0.56	0.06	0.05	0.46	0.89	0.74	0.52	0.5	0.55
Kidney	0.95	0.61	0.06	0.07	0.23	0.95	0.94	0.88	0.86	0.8
Liver	0.9	0.33	0.04	0.04	0.57	0.88	0.38	0.23	0.22	0.59
Lung	0.85	0.59	0.15	0.18	0.42	0.81	0.83	0.69	0.65	0.61
Ovary	0.78	0.37	0.01	0.01	0.21	0.78	0.61	0.42	0.4	0.43
Pancreas	0.79	0.48	0.11	0.09	0.44	0.79	0.66	0.31	0.3	0.54
Prostate	0.84	0.56	0.1	0.09	0.47	0.81	0.8	0.56	0.52	0.58
Skin	0.95	0.77	0.58	0.55	0.71	0.96	0.93	0.8	0.77	0.79
Stomach	0.82	0.52	0.23	0.21	0.51	0.74	0.62	0.59	0.57	0.68

Figure 12 Comparative assessment on pan-cancer patients - correlation per tissue type breakdown. Each entry represents the mean correlation value between the predicted exposures and the COSMIC exposures over 10 repetitions of the experiment. In each experiment, 30 mutations are randomly sampled per patient and the relative signature exposures are predicted using each method. The same 30 mutations are used to evaluate all methods in each repetition.

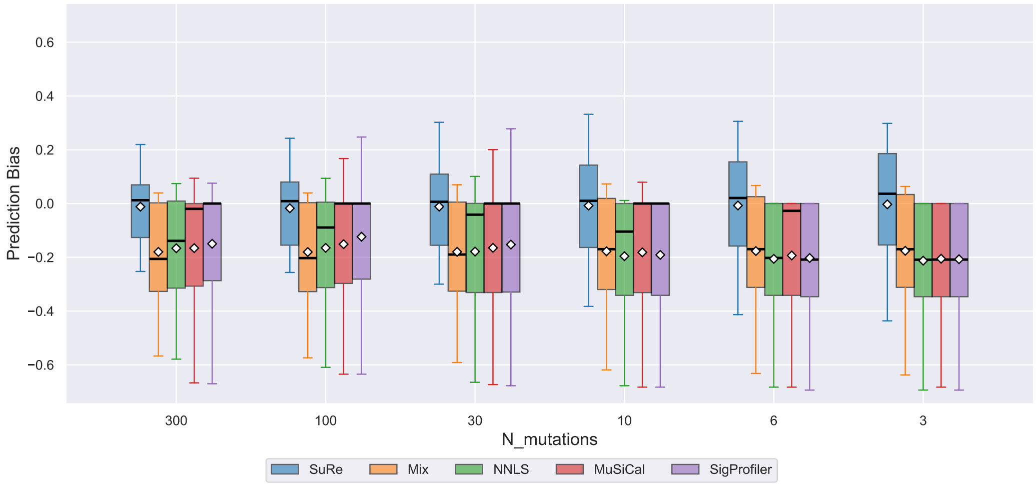


Figure 13 Prediction bias analysis for SBS3 prediction in breast cancer. Each box represents the values over $\lceil\frac{300}{m}\rceil$ repetitions, where m is the number mutations per sample. The same m mutations are used to evaluate all methods in each repetition. The black line within each whisker represents to the median value and the diamond marker represents the mean value. Whiskers span from the 10th to the 90th percentile.

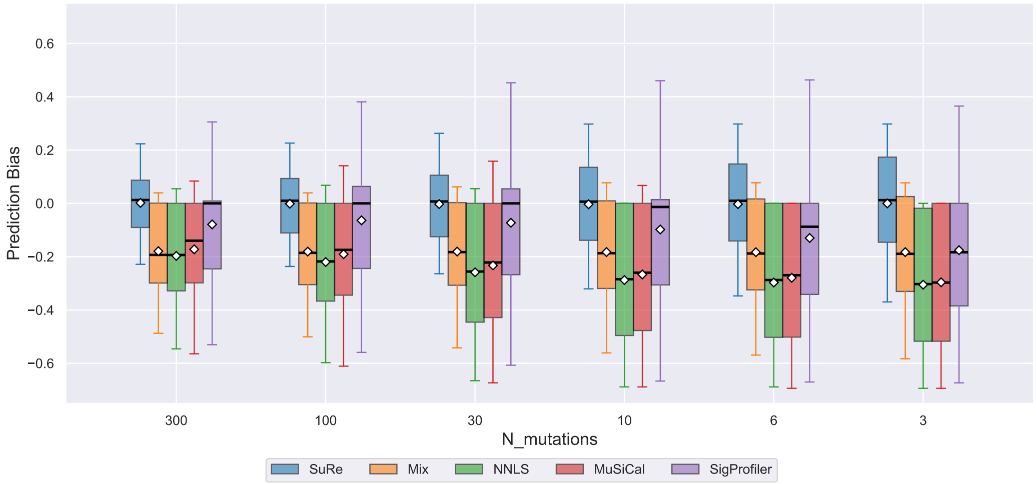


Figure 14 Prediction bias analysis for flat signatures that are common in breast cancer (SBS3 and SBS5). Each box represents the values over $\lceil\frac{300}{m}\rceil$ repetitions, where m is the number mutations per sample. The same m mutations are used to evaluate all methods in each repetition. The black line within each whisker represents to the median value and the diamond marker represents the mean value. Whiskers span from the 10th to the 90th percentile.

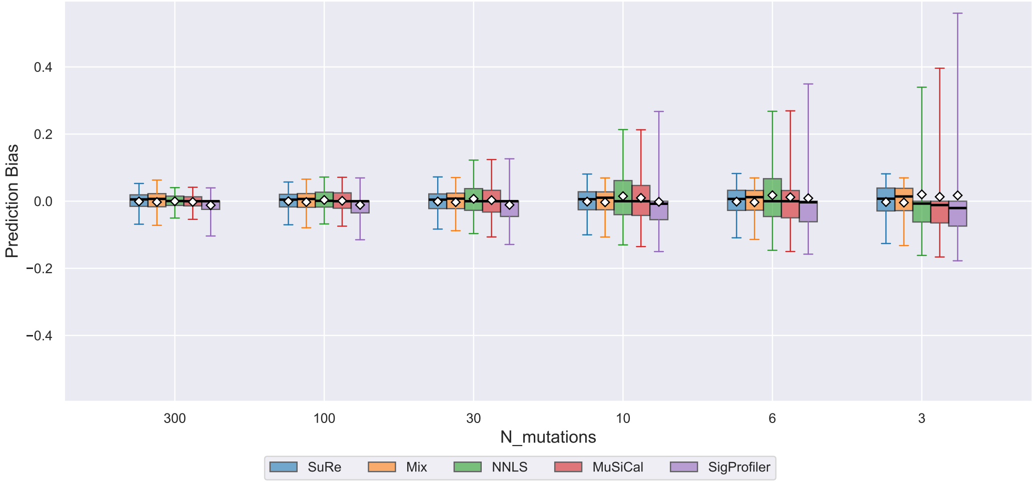


Figure 15 Prediction bias analysis for spiky signatures that are common in breast cancer (SBS1, SBS2, SBS13, SBS18, SBS34). Each box represents the values over $\lceil \frac{300}{m} \rceil$ repetitions, where m is the number mutations per sample. The same m mutations are used to evaluate all methods in each repetition. The black line within each whisker represents to the median value and the diamond marker represents the mean value. Whiskers span from the 10th to the 90th percentile.

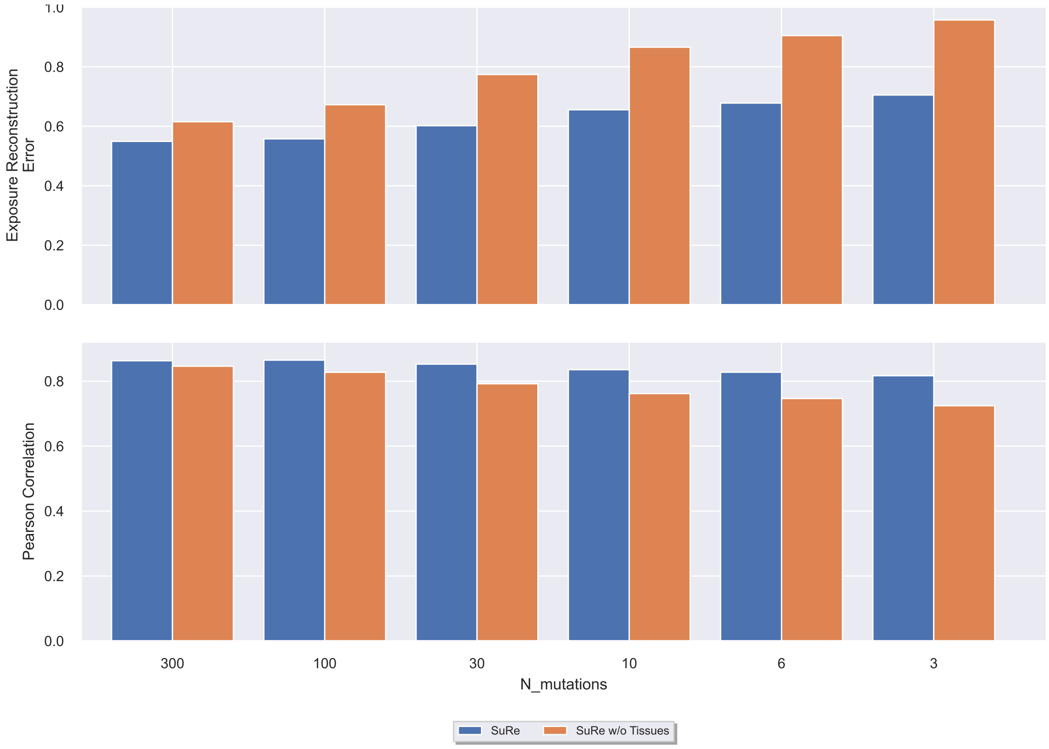


Figure 16 Comparative assessment on pan-cancer patients - SuRe vs SuRe without tissue label data. Each bar represents the mean value over $\lceil \frac{300}{m} \rceil$ repetitions, where m is the number mutations per sample. The same m mutations are used to evaluate all methods in each repetition.