Algorithms for Molecular Biology https://doi.org/10.1186/s13015-026-00305-0

Article in Press

Mutational signature refitting on sparse pan-cancer data

Received: 25 Nov 2025

Accepted: 30 Jun 2026

Published online: 17 July 2026

Cite this article as: Gilad, G., Przytycka, T., Sharan, R. *et al.* Mutational signature refitting on sparse pan-cancer data. *Algorithms Mol Biol* (2026). https://doi.org/10.1186/s13015-026-00305-0

Gal Gilad, Teresa Przytycka & Roded Sharan

Mutational signature refitting on sparse pan-cancer data

Gal Gilad[1], Teresa M. Przytycka[2], Roded Sharan[1*]

[1*]Blavatnik School of Computer Science and AI, Tel Aviv University, Tel Aviv, 69978, Israel.
[2]National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, 20892, MD, USA.

*Corresponding author(s). E-mail(s): roded@tauex.tau.ac.il;

Abstract

Background: 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 accurately decipher those exposures. Previous work in this domain was unsupervised in nature, employing algebraic decomposition and probabilistic inference methods.
Results: We present SuRe, a supervised approach to signature refitting that demonstrates superiority over current methods. 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 both tumor-type-specific and pan-cancer data sets, with an increasing performance advantage as the data become sparser.
Conclusion: We further demonstrate the model's utility in clinical settings by predicting homologous recombination deficiency in breast cancer from sparse data. Furthermore, SuRe outperforms standard methods in the unsupervised stratification of over 13,000 patients from large-scale panel sequencing cohorts, highlighting its potential for analyzing targeted sequencing data.

Keywords: Mutational signatures, Signature refitting, cancer genomics, genomic data analysis, somatic mutations

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 [1, 2]. These signatures are cataloged in the COSMIC database [3].

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) [4]. 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 [5], 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 [6], a heuristic based on the iterative application of the multiple linear regression and exclusion of signatures with low relative exposures; SigLASSO [7], which seeks to produce sparse, high-confidence solutions by jointly optimizing L1 regularized signature refitting and mutation sampling likelihood; SignatureEstimation [8] that evaluates the stability and confidence of signature activity levels by applying perturbations to the mutation count inputs; SigNet [9], 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) [1, 10], 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 [11], 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 [12].

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 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.

2.2.1 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.

Although standard preprocessing often involves converting mutation counts into compositional probabilities, SuRe is explicitly designed to process raw, absolute mutation counts. This design choice is biologically motivated: an increased overall mutation burden is typically driven by an additive surge in a specific mutational process (e.g., mismatch repair deficiency) rather than a uniform, proportional scaling of all background signatures. Preserving absolute counts ensures this additive signal is not obscured, whereas converting counts to a relative composition masks the absolute stability of the underlying background processes. Computationally, retaining raw counts allows the network's pre-activation logit magnitudes to scale naturally with the volume of available evidence. In the context of highly sparse panel data, this ensures the model outputs more conservative (higher entropy) probability distributions, while producing sharper, higher-confidence predictions when the absolute mutational signal is strong. To ensure stable convergence and mitigate potential instability across vastly different input scales, SuRe dynamically subsamples the number of mutations (m) during training, explicitly forcing the network to learn robust representations across the full continuum of mutation burdens.

The model's loss is the Kullback-Leibler (KL) divergence between the predicted relative exposure vector $\hat{y}_i$ and the actual relative exposure vector y_i for sample i

(summed over all samples). We opted for KL divergence rather than a multinomial likelihood loss due to two fundamental constraints. First, the COSMIC database provides aggregate exposures per sample rather than discrete ground-truth labels for individual mutations, precluding a discrete likelihood approach against the target exposures. Second, computing a likelihood against the highly sparse and noisy observed input counts would revert the framework to an unsupervised reconstruction objective, incentivizing the network to overfit to sampling noise. Thus, utilizing KL divergence against the aggregate WGS exposures provides a stable target distribution that avoids overfitting to sparse inputs.

The model is based on a Mixture-of-Experts [13] 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 $96 + T$ neurons. 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 [14], using a Stochastic Gradient Descent (SGD) optimizer and a cosine annealing with warm restarts [15] 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.

2.2.2 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,

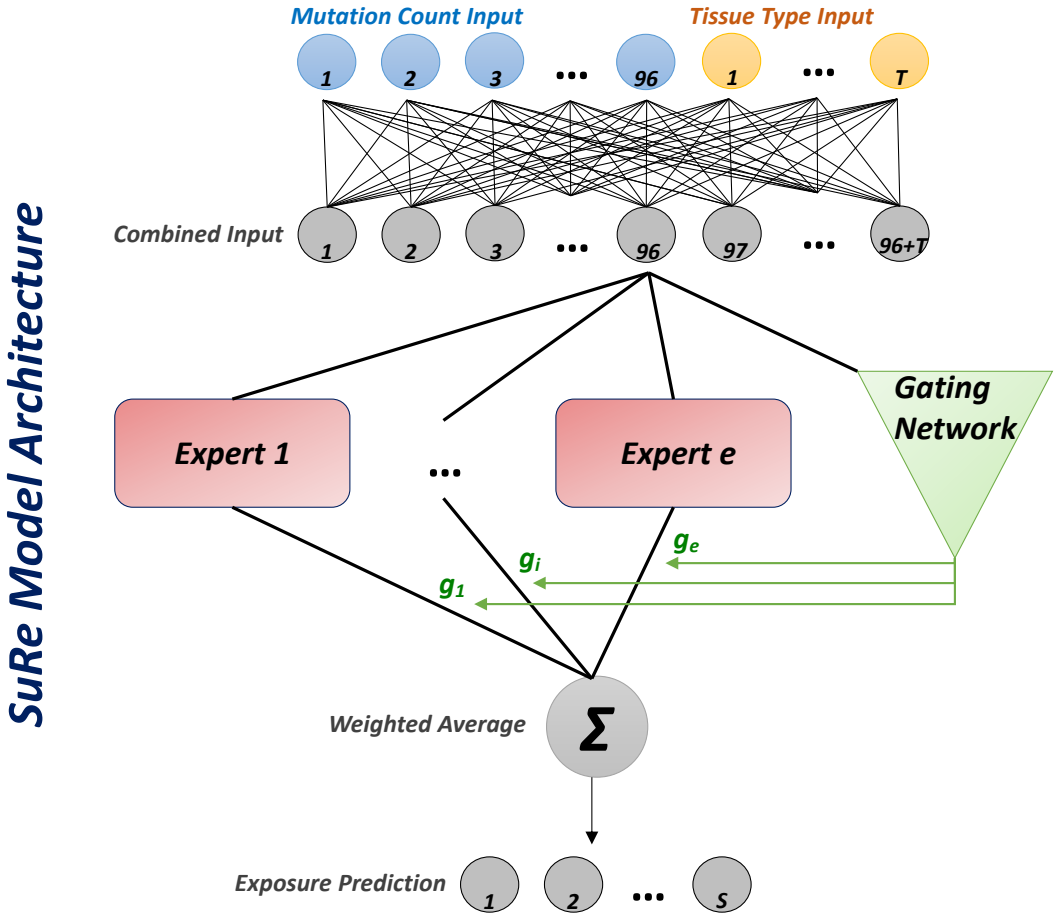

Fig. 1 SuRe Mixture-of-Experts architecture. The tissue type input (size T) is concatenated to the mutation count input (96) and fed into a layer of the same size (96+T). This combined representation of the input is fed into e expert modules, as well as a gating network. Each expert outputs probabilities 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.

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. While performing a separate, exhaustive search for the learning rate on the pan-cancer data set could potentially yield marginal improvements, we adopted this heuristic approach due to the significant computational cost of full hyperparameter tuning on the much larger cohort. As demonstrated in our pan-cancer evaluations, this configuration still achieved robust convergence and outperformed existing methods. 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 $h = 500$.

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.

2.3 Architecture Validation and Expert Probing

To validate our choice of a Mixture-of-Experts (MoE) architecture, we performed two analyses on the pan-cancer model, with full results available in Supplementary Section S1.

First, we conducted an ablation study comparing our MoE model against a baseline standard fully connected network with a similar number of parameters. Both models were trained and evaluated on the pan-cancer dataset across multiple levels of data sparsity to assess the architectural advantage on heterogeneous data.

Second, to understand the MoE's internal strategy, we performed an expert probing analysis on a representative sparse dataset (m=30 mutations). We computed the Pearson correlation between each expert's gating weight and the relative exposure of each mutational signature, thereby identifying any learned expert specialization.

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.

2.5 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 [1], and ii) re-analysis of the data using a newer version of the software, published in [10]. 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 [10] and [16] 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 which consists of single base substitution patterns of 2,700 simulated cancer genomes, corresponding to 300 tumors from nine different cancer types, generated using 21 different COSMIC reference signatures. 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. [17], along with their HRD status labels, predicted by HRDetect [18]. 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 [19] panel and binarized the labels by excluding the

13 “intermediate” labeled samples, leaving 224 samples, 62% of them with predicted HRD.

2.6 Clinical stratification

To evaluate performance on native sparse data, we analyzed 13,076 samples from the MSK-IMPACT [20] and MSK-MET cohorts [21]. We focused our analysis on the three largest tissue cohorts available in both datasets that were also represented in our pan-cancer training data: lung ($n = 6,666$), breast ($n = 3,833$), and prostate ($n = 2,577$).

We selected clinical attributes known to associate with specific mutational processes to serve as biological benchmarks. We evaluated the association with metastatic disease (comparing primary vs. metastatic samples) across all three tissue types. Additionally, for lung cancer, we utilized smoking history data (available in the MSK-IMPACT cohort), distinguishing "Previous/Current Smokers" from "Never Smokers" while excluding "Unknown" status. For breast and prostate cancers, we evaluated the association with Homologous Recombination (HR) pathway alterations. The latter was defined by the presence of somatic non-synonymous mutations in *BRCA1*, *BRCA2*, or *ATM*. Finally, we also tested for association with Tumor Mutational Burden (TMB) status (High vs. Low, defined by the cohort median) to assess whether the inferred signature profiles capture the distinct mutational patterns associated with high tumor burden.

For each analysis, mutation counts were refitted using SuRe, NNLS, and SigProfiler. We restricted our comparison to these two methods as they represent the most widely established benchmarks for mutational signature analysis in clinical and genomic studies. To ensure that stratification was driven by mutational composition rather than burden, the predicted exposure vectors were L1-normalized to sum to 1. We then applied K-Means clustering ($k = 2$) to the normalized exposures for each tissue and dataset separately, and evaluated the enrichment of the clinical labels within the resulting clusters using Fisher's Exact Test.

3 Results

To evaluate the performance of SuRe and demonstrate its clinical applications, we conducted comprehensive testing on both synthetic and real data sets. To assess our algorithm's robustness under varying degrees of data sparsity, we constructed simulated data sets by downsampling the original whole-genome sequencing (WGS) data, 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: SigProfilerAssignment 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 Model Rationale and Architecture Validation

We first validated our choice of a Mixture-of-Experts (MoE) architecture for handling complex, heterogeneous pan-cancer data. An ablation study confirmed that while a simpler, standard fully-connected network performs competitively, our MoE architecture provides a clear and consistent performance advantage (see Supplementary Figure S1).

To understand this advantage, we analyzed the model's internal mechanics on sparse data ($m = 30$). The analysis revealed that the MoE learns to function as an interpretable team of specialists, with different experts developing strong correlations to specific signatures (e.g., common clock-like signatures, UV signatures, APOBEC signatures, etc.) (see Supplementary Figure S2). A full analysis is provided in Supplementary Section S1.

3.2 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 [10] and refitting [16] 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. To quantify this advantage, we compared SuRe's performance against the second-best method using a Wilcoxon signed-rank test. Due to the $\lceil \frac{300}{m} \rceil$ downsampling repetition strategy, sample sizes at $m = 300$ ($N = 1$) and $m = 100$ ($N = 3$) are mathematically insufficient for non-parametric significance testing. However, in the sparse data regimes that are the primary focus of this study ($m \leq 30$, $N \geq 10$), SuRe demonstrated a statistically significant advantage, outperforming the next best method across all repetitions. Supplementary Figure S3 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.

3.3 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 inferring the underlying mutation distribution. In this analysis, we used only mutational signatures that were active in at least

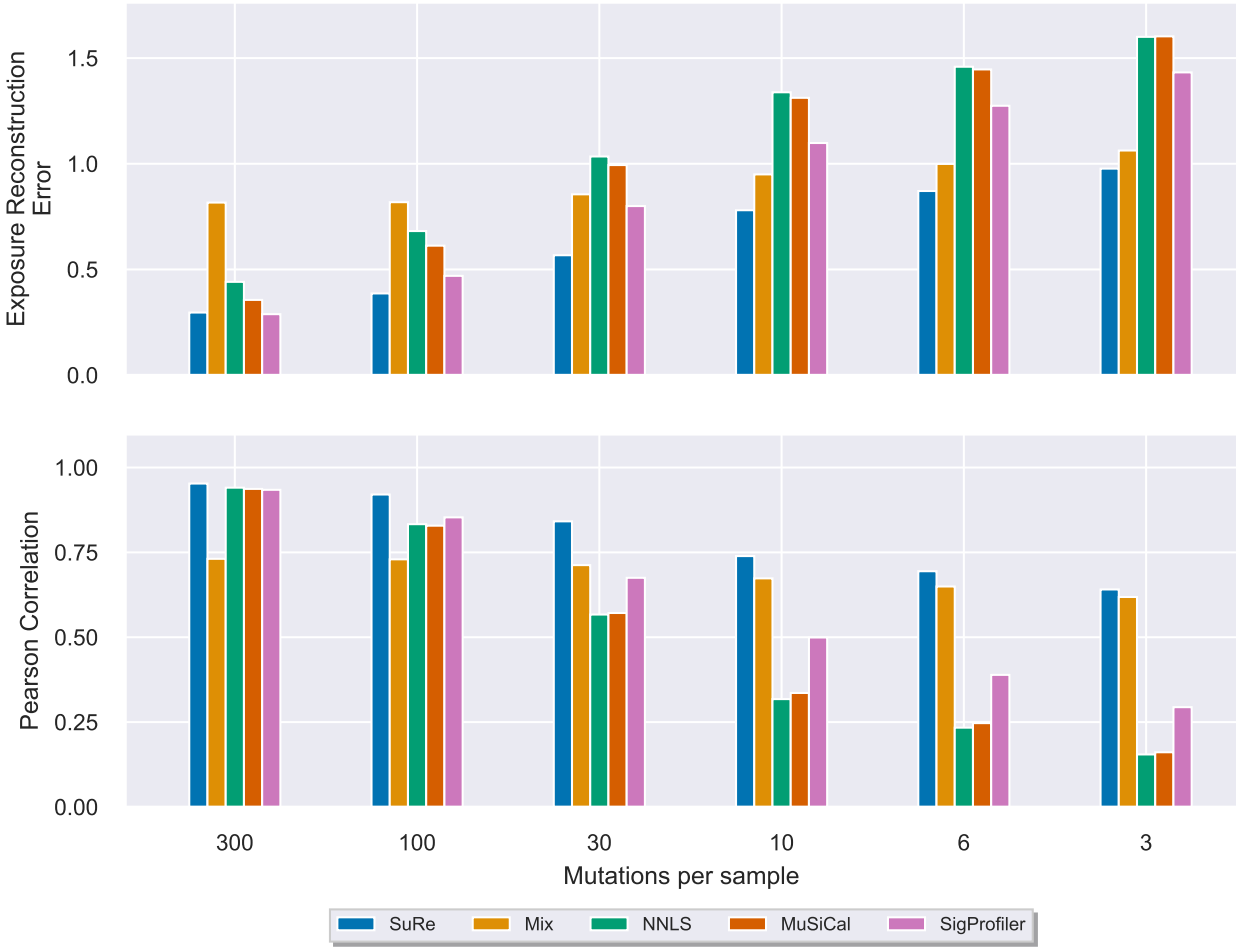

Fig. 2 Performance evaluation on synthetic data. Each bar represents the mean value over $\lceil \frac{300}{m} \rceil$ repetitions, where m is the number mutations per sample. The same m mutations are used to evaluate all methods in each repetition.

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 S4, and

again demonstrate that SuRe is more resilient to the challenges of refitting to a large database of potentially inactive signatures.

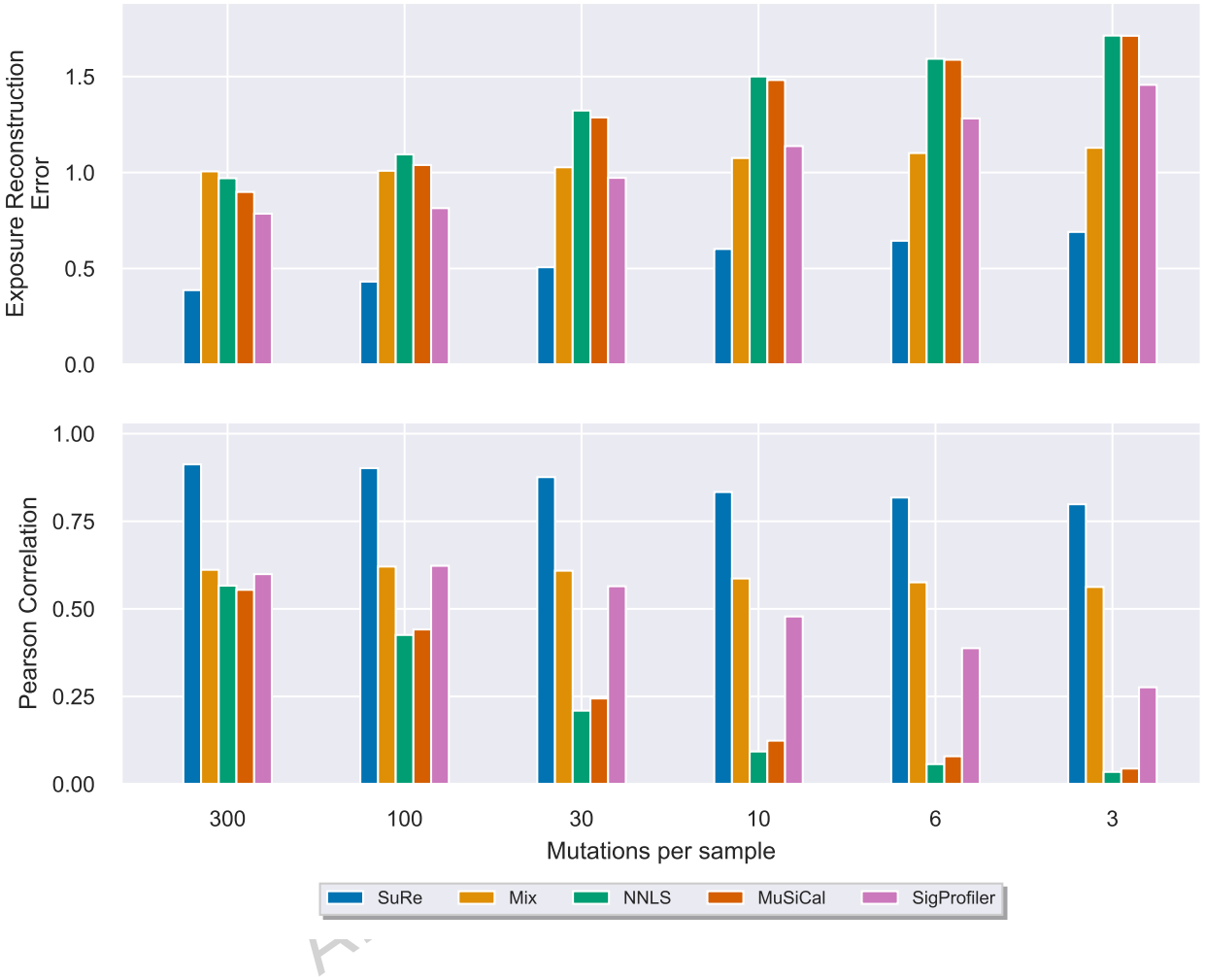

Fig. 3 Comparative assessment of exposure prediction in breast cancer patients. Each value represents the mean over $\lceil \frac{300}{m} \rceil$ repetitions, where m is the number mutations per sample. The same m mutations are used to evaluate all methods in each repetition.

More recently, a re-analysis of COSMIC samples was published based on a new and improved version of the SigProfiler software [10]. 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 S5 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 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 S6).

3.4 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 [11]. 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 S7. To assess these results in a more specific way, we focus on a mid-range downsampling size of 30 mutations for the following analyses (Supplementary Figures S8 and S9). 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 S13.

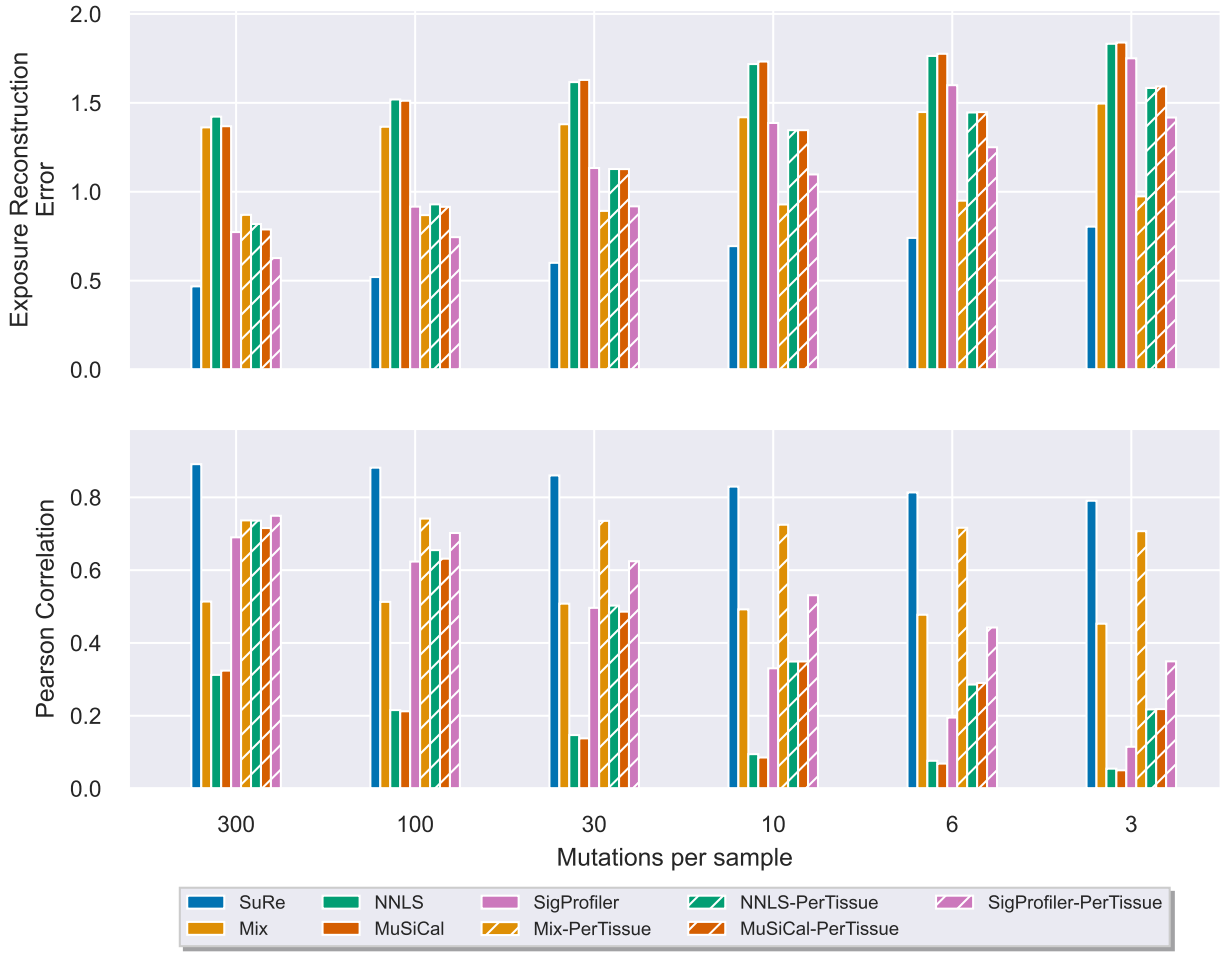

Fig. 4 Comparative assessment - targeting COSMIC exposures on pan-cancer patients. Each bar represents the mean value over $\lceil \frac{300}{m} \rceil$ repetitions, where m is the number mutations per sample. The same m mutations are used to evaluate all methods in each repetition.

3.5 Prediction of HRD status

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 [17], we downsampled the mutation data to the genomic regions of the MSK-IMPACT panel [19]. 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 [17], we used the inferred SBS3 exposure to estimate HRD. That is, the predicted absolute SBS3 exposures (calculated as the inferred relative exposure multiplied by the total mutation burden) 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 S10). 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 S11 and S12).

3.6 Stratification of panel data

Next, we sought to validate SuRe on native sparse data, rather than downsampled WGS. We applied SuRe, NNLS, and SigProfiler to 13,076 samples across Lung ($n = 6,666$), Breast ($n = 3,833$), and Prostate ($n = 2,577$) cancers from the MSK-IMPACT and MSK-MET cohorts. We performed clustering of patients based on their predicted relative signature exposures to test if the methods could recover clinically relevant subtypes.

We first evaluated the ability of the methods to stratify patients based on their tumor mutational burden (TMB) (Figure 6). Since the input exposures were normalized prior to clustering, any stratification here indicates the detection of distinct mutational compositions (e.g., mismatch repair deficiency signatures) associated with hypermutation, rather than a proxy of the mutation count itself. SuRe produced clusters with extremely strong associations to high TMB across all tissue types ($p < 10^{-40}$ in lung and breast). Most notably, in the prostate cancer cohort of the MSK-IMPACT dataset, a tissue with typically lower mutation counts, SuRe was the only method to identify a significant TMB-associated subgroup ($p < 10^{-16}$), while both NNLS

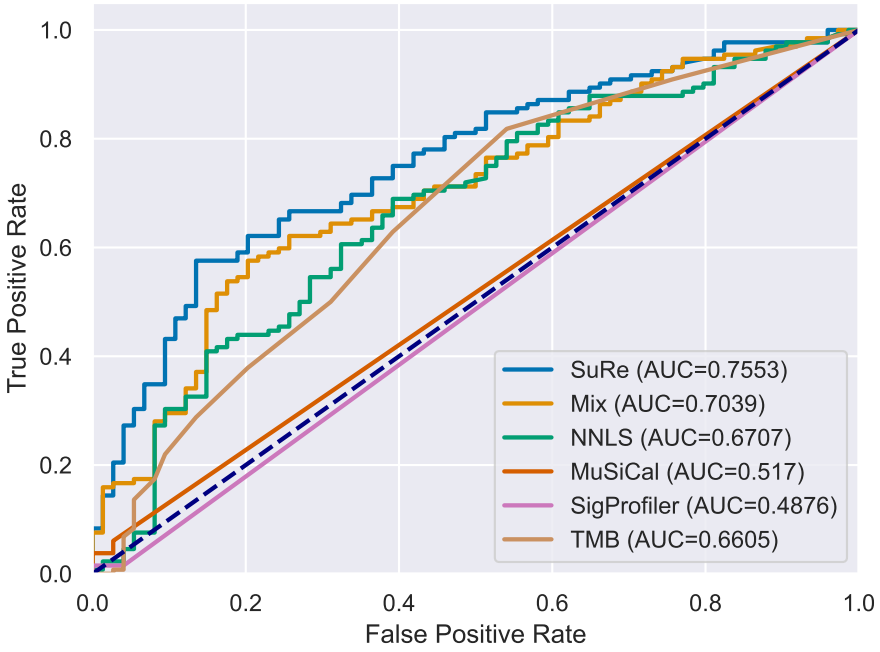

Fig. 5 ROC curves for HR deficiency prediction - SBS3-optimized breast cancer model. The curves are generated by varying the decision threshold on the predicted absolute exposure of signature SBS3 (calculated as the inferred relative exposure multiplied by the total mutation burden) to classify samples as HRD positive or negative.

($p = 0.16$) and SigProfiler ($p = 0.37$) failed to stratify the patients into statistically significant groups.

Regarding clinical outcomes, SuRe consistently produced clusters with tighter associations to metastatic status than competing methods. In the MSK-IMPACT prostate cancer cohort, SuRe was the only method to identify a signature-based cluster significantly enriched for metastatic samples ($p = 0.032$), whereas NNLS ($p = 0.57$) and SigProfiler ($p = 0.18$) did not recover this signal. Similarly, in the breast cancer cohort of the MSK-MET dataset, SuRe strongly stratified primary from metastatic samples ($p < 10^{-12}$), significantly outperforming SigProfiler ($p = 0.027$) and NNLS ($p = 0.10$). We also tested for associations with metastatic status in lung cancer, but no method produced significant stratification in either dataset ($p > 0.05$).

We next examined genotypic associations. In breast cancer samples (MSK-IMPACT and MSK-MET), SuRe accurately recovered a subgroup enriched for HR pathway alterations ($p < 10^{-11}$ in both datasets), consistently outperforming NNLS ($p \approx 10^{-8}$ and $p = 0.54$) and SigProfiler ($p > 0.009$). Associations with HR pathway alterations in prostate cancer were found to be insignificant for all methods.

In lung cancer samples from the MSK-IMPACT cohort (where smoking history was available), all methods successfully stratified patients by smoking status. However,

SuRe provided a clearer separation of the clinical phenotypes: it defined a cluster where 91% of patients were smokers (compared to 74% in the remaining cluster), yielding a separation gap of 17%, compared to a 10% gap for SigProfiler, without a clearly defined smoker-associated group.

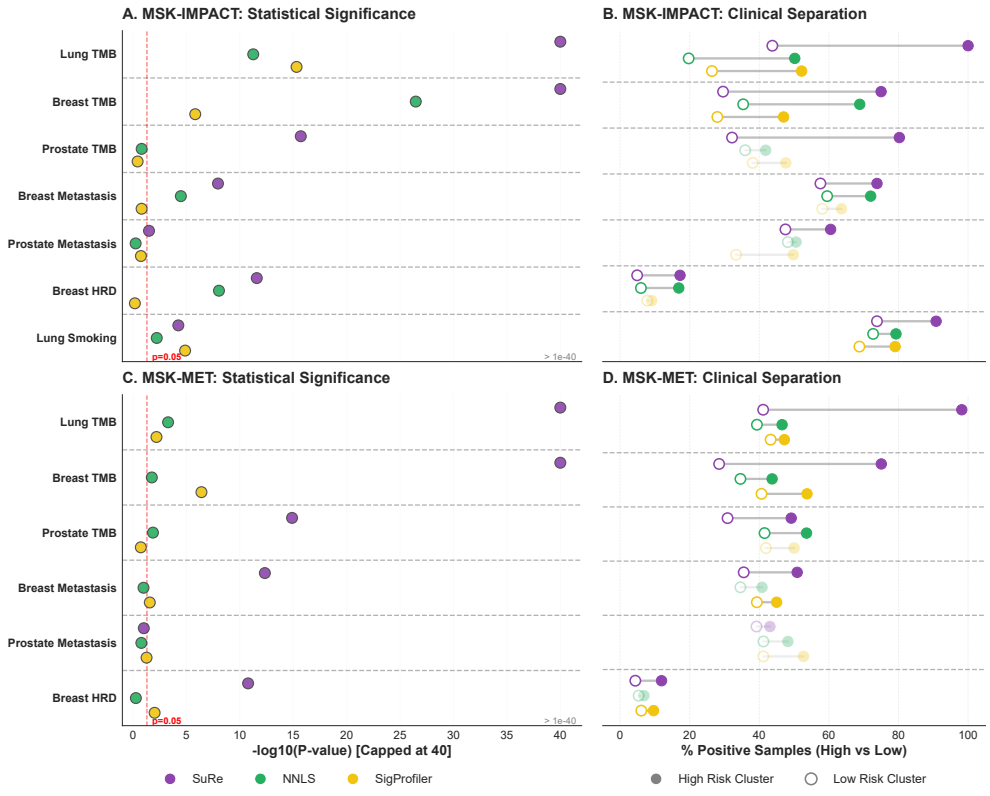

Fig. 6 Stratification of panel sequencing cohorts. (A, C) Statistical significance (-log10 p-value) of the association between unsupervised clusters (k=2) of predicted exposures and clinical labels in MSK-IMPACT and MSK-MET datasets. SuRe (purple) consistently achieves higher significance than NNLS (green) and SigProfiler (yellow). (B, D) Dumbbell plots showing the separation power. The dots represent the percentage of positive cases (e.g., Smokers, Metastatic) in the two identified clusters. A wider gap indicates better separation of the clinical phenotype based on relative signature exposures alone. Note: Associations for Lung Metastasis and Prostate HRD were tested but found to be insignificant for all methods (not shown).

4 Conclusions

We have presented SuRe, a supervised method for mutational signature refitting on sparse mutation data. By capturing correlations between signatures, our model outperforms state-of-the-art unsupervised methods in reconstructing exposure profiles from limited mutation counts, particularly as data sparsity increases. We further showed that SuRe significantly improves the detection of homologous recombination deficiency from targeted sequencing data compared to established methods, demonstrating its ability to recover complex biological signals where current approaches struggle.

Our unsupervised analysis of over 13,000 samples from the MSK-IMPACT and MSK-MET cohorts highlights the broader utility of SuRe for the retrospective mining of large panel databases. While signature inference on sparse data is often confounded by Tumor Mutational Burden (TMB), SuRe leverages normalized signature composition to robustly stratify patients. This approach allowed the model to identify clinical subtypes in settings where standard methods failed to recover a significant signal, such as identifying metastatic subgroups in prostate cancer, a tissue characterized by low mutation counts. This suggests that SuRe can extract biologically relevant mutational patterns even when restricted to sparse input 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 [1, 10]. 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 [1, 22], tobacco smoking history [23], homologous recombination deficiency status [18] and mismatch repair deficiency [22]. 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.

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 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.

Supplementary information. Additional file 1: Supplementary Material. Contains Supplementary Section 1 (Model Architecture Validation and Expert Probing; Supplementary Figures S1–S2) and Supplementary Section 2 (Extended Performance Evaluation and Bias Analysis; Supplementary Figures S3–S13).

Declarations

- **Funding**: This research was supported in part by grants from the US-Israel Binational Science Foundation and the Alrov Center for Digital Medicine, and in part by the Division of Intramural Research of the National Library of Medicine (NLM), National Institutes of Health (NIH). The contributions of the NIH author are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

- **Conflict of interest**: The authors declare that they have no competing interests.
- **Data Ethics approval and consent to participate**: Not applicable.
- **Consent for publication**: Not applicable.
- **Data availability**: The whole genome sequencing mutation catalogs used in this study are available at Synapse (https://www.synapse.org/#!Synapse:syn11726616) [1] and Figshare (https://doi.org/10.6084/m9.figshare.20409430) [10]. The corresponding mutational signature exposures are available at Synapse (https://www.synapse.org/#!Synapse:syn11804040) [1]. The triple-negative breast cancer dataset is available from Staaf et al. [17]. Data from the MSK-IMPACT and MSK-MET cohorts are publicly available through cBioPortal and were originally published by Zehir et al. [20] and Nguyen et al. [21].
- **Materials availability**: Not applicable.
- **Code availability**: The SuRe source code is available at https://github.com/GalGilad/SuRe. A frozen version of the code used to generate the results in this study is archived on Zenodo at https://doi.org/10.5281/zenodo.20185762.

References

[1] Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., Islam, S.M.A., Lopez-Bigas, N., Klimczak, L.J., McPherson, J.R., Morganella, S., Sabarinathan, R., Wheeler, D.A., Mustonen, V., Boutros, P., Chan, K., Fujimoto, A., Getz, G., Kazanov, M., Lawrence, M., Martincorena, I., Nakagawa, H., Polak, P., Prokopec, S., Roberts, S.A., Rozen, S.G., Saini, N., Shibata, T., Shiraishi, Y., Stratton, M.R., Teh, B.T., Vázquez-García, I., Yousif, F., Yu, W., Aaltonen, L.A., Abascal, F., Abeshouse, A., Aburatani, H., Adams, D.J., Agrawal, N., Ahn, K.S., Ahn, S.-M., Aikata, H., Akbani, R., Akdemir, K.C., Al-Ahmadie, H., Al-Sedairy, S.T., Al-Shahrour, F., Alawi, M., Albert, M., Aldape, K., Ally, A., Alsop, K., Alvarez, E.G., Amary, F., Amin, S.B., Aminou, B., Ammerpohl, O., Anderson, M.J., Ang, Y., Antonello, D., Anur, P., Aparicio, S., Appelbaum, E.L., Arai, Y., Aretz, A., Arihiro, K., Ariizumi, S.-i., Armenia, J., Arnould, L., Asa, S., Assenov, Y., Atwal, G., Aukema, S., Auman, J.T., Aure, M.R.R., Awadalla, P., Aymerich, M., Bader, G.D., Baez-Ortega, A., Bailey, M.H., Bailey, P.J., Balasundaram, M., Balu, S., Bandopadhayay, P., Banks, R.E., Barbi, S., Barbour, A.P., Barenboim, J., Barnholtz-Sloan, J., Barr, H., Barrera, E., Bartlett, J., Bartolome, J., Bassi, C., Bathe, O.F., Baumhoer, D., Bavi, P., Baylin, S.B., Bazant,

W., Beardsmore, D., Beck, T.A., Behjati, S., Behren, A., Niu, B., Bell, C., Beltran, S., Benz, C., Berchuck, A., Bergmann, A.K., Berman, B.P., Berney, D.M., Bernhart, S.H., Beroukhim, R., Berrios, M., Bersani, S., Bertl, J., Betancourt, M., Bhandari, V., Bhosle, S.G., Biankin, A.V., Bieg, M., Bigner, D., Binder, H., Birney, E., Birrer, M., Biswas, N.K., Bjerkehagen, B., Bodenheimer, T., Boice, L., Bonizzato, G., De Bono, J.S., Bootwalla, M.S., Borg, A., Borkhardt, A., Boroevich, K.A., Borozan, I., Borst, C., Bosenberg, M., Bosio, M., Boultwood, J., Bourque, G., Boutros, P.C., Bova, G.S., Bowen, D.T., Bowlby, R., Bowtell, D.D.L., Boyault, S., Boyce, R., Boyd, J., Brazma, A., Brennan, P., Brewer, D.S., Brinkman, A.B., Bristow, R.G., Broaddus, R.R., Brock, J.E., Brock, M., Broeks, A., Brooks, A.N., Brooks, D., Brors, B., Brunak, S., Bruxner, T.J.C., Bruzos, A.L., Buchanan, A., Buchhalter, I., Buchholz, C., Bullman, S., Burke, H., Burkhardt, B., Burns, K.H., Busanovich, J., Bustamante, C.D., Butler, A.P., Butte, A.J., Byrne, N.J., Børresen-Dale, A.-L., Caesar-Johnson, S.J., Cafferkey, A., Cahill, D., Calabrese, C., Caldas, C., Calvo, F., Camacho, N., Campbell, P.J., Campo, E., Cantù, C., Cao, S., Carey, T.E., Carlevaro-Fita, J., Carlsen, R., Cataldo, I., Cazzola, M., Cebon, J., Cerfolio, R., Chadwick, D.E., Chakravarty, D., Chalmers, D., Chan, C.W.Y., Chan-Seng-Yue, M., Chandan, V.S., Chang, D.K., Chanock, S.J., Chantrill, L.A., Chateigner, A., Chatterjee, N., Chayama, K., Chen, H.-W., Chen, J., Chen, K., Chen, Y., Chen, Z., Cherniack, A.D., Chien, J., Chiew, Y.-E., Chin, S.-F., Cho, J., Cho, S., Choi, J.K., Choi, W., Chomienne, C., Chong, Z., Choo, S.P., Chou, A., Christ, A.N., Christie, E.L., Chuah, E., Cibulskis, C., Cibulskis, K., Cingarlini, S., Clapham, P., Claviez, A., Cleary, S., Cloonan, N., Cmero, M., Collins, C.C., Connor, A.A., Cooke, S.L., Cooper, C.S., Cope, L., Corbo, V., Cordes, M.G., Cordner, S.M., Cortés-Ciriano, I., Covington, K., Cowin, P.A., Craft, B., Craft, D., Creighton, C.J., Cun, Y., Curley, E., Cutcutache, I., Czajka, K., Czerniak, B., Dagg, R.A., Danilova, L., Davi, M.V., Davidson, N.R., Davies, H., Davis, I.J., Davis-Dusenbery, B.N., Dawson, K.J., De La Vega, F.M., De Paoli-Iseppi, R., Defreitas, T., Tos, A.P.D., Delaneau, O., Demchok, J.A., Group, P.M.S.W., Consortium, P.C.A.W.G.: The repertoire of mutational signatures in human cancer. Nature **578**(7793), 94–101 (2020) https://doi.org/10.1038/s41586-020-1943-3

[2] Degasperi, A., Zou, X., Amarante, T.D., Martinez-Martinez, A., Koh, G.C.C., Dias, J.M.L., Heskin, L., Chmelova, L., Rinaldi, G., Wang, V.Y.W., Nanda, A.S., Bernstein, A., Momen, S.E., Young, J., Perez-Gil, D., Memari, Y., Badja, C., Shooter, S., Czarnecki, J., Brown, M.A., Davies, H.R., null, Nik-Zainal, S., Ambrose, J.C., Arumugam, P., Bevers, R., Bleda, M., Boardman-Pretty, F., Boustred, C.R., Brittain, H., Caulfield, M.J., Chan, G.C., Fowler, T., Giess, A., Hamblin, A., Henderson, S., Hubbard, T.J.P., Jackson, R., Jones, L.J., Kasperaviciute, D., Kayikci, M., Kousathanas, A., Lahnstein, L., Leigh, S.E.A., Leong, I.U.S., Lopez, F.J., Maleady-Crowe, F., McEntagart, M., Minneci, F., Moutsianas, L., Mueller, M., Murugaesu, N., Need, A.C., O'Donovan, P., Odhams, C.A., Patch, C., Perez-Gil, D., Pereira, M.B., Pullinger, J., Rahim, T., Rendon, A., Rogers, T., Savage, K., Sawant, K., Scott, R.H., Siddiq, A., Sieghart, A., Smith, S.C., Sosinsky, A., Stuckey, A., Tanguy, M., Tavares, A.L.T., Thomas,

E.R.A., Thompson, S.R., Tucci, A., Welland, M.J., Williams, E., Witkowska, K., Wood, S.M.: Substitution mutational signatures in whole-genome-sequenced cancers in the uk population. Science **376**(6591), 9283 (2022) https://doi.org/10.1126/science.abl9283 https://www.science.org/doi/pdf/10.1126/science.abl9283

[3] Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S.C., Kok, C.Y., Noble, K., Ponting, L., Ramshaw, C.C., Rye, C.E., Speedy, H.E., Stefancsik, R., Thompson, S.L., Wang, S., Ward, S., Campbell, P.J., Forbes, S.A.: COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Research **47**(D1), 941–947 (2018) https://doi.org/10.1093/nar/gky1015 https://academic.oup.com/nar/article-pdf/47/D1/D941/27441712/gky1015.pdf

[4] Gentleman, W.M.: Solving least squares problems (charles l. lawson and richard j. hanson). SIAM Review **18**(3), 518–520 (1976) https://doi.org/10.1137/1018100 https://doi.org/10.1137/1018100

[5] Sason, I., Chen, Y., Leiserson, M.D.M., Sharan, R.: A mixture model for signature discovery from sparse mutation data. Genome Medicine **13**(1), 173 (2021) https://doi.org/10.1186/s13073-021-00988-7

[6] Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S., Swanton, C.: deconstructsigs: delineating mutational processes in single tumors distinguishes dna repair deficiencies and patterns of carcinoma evolution. Genome Biology **17**(1), 31 (2016) https://doi.org/10.1186/s13059-016-0893-4

[7] Li, S., Crawford, F.W., Gerstein, M.B.: Using siglasso to optimize cancer mutation signatures jointly with sampling likelihood. Nature Communications **11**(1), 3575 (2020) https://doi.org/10.1038/s41467-020-17388-x

[8] Huang, X., Wojtowicz, D., Przytycka, T.M.: Detecting presence of mutational signatures in cancer with confidence. Bioinformatics **34**(2), 330–337 (2017) https://doi.org/10.1093/bioinformatics/btx604 https://academic.oup.com/bioinformatics/article-pdf/34/2/330/48913076/bioinformatics_34_2_330_s1.pdf

[9] Serrano Colome, C., Canal Anton, O., Seplyarskiy, V., Weghorn, D.: Mutational signature decomposition with deep neural networks reveals origins of clock-like processes and hypoxia dependencies. bioRxiv (2023) https://doi.org/10.1101/2023.12.06.570467 https://www.biorxiv.org/content/early/2023/12/08/2023.12.06.570467.full.pdf

[10] Islam, S.M.A., Díaz-Gay, M., Wu, Y., Barnes, M., Vangara, R., Bergstrom, E.N., He, Y., Vella, M., Wang, J., Teague, J.W., Clapham, P., Moody, S., Senkin, S., Li, Y.R., Riva, L., Zhang, T., Gruber, A.J., Steele, C.D., Otlu, B., Khandekar, A., Abbasi, A., Humphreys, L., Syulyukina, N., Brady, S.W., Alexandrov, B.S., Pillay, N., Zhang, J., Adams, D.J., Martincorena, I., Wedge, D.C., Landi, M.T., Brennan,

P., Stratton, M.R., Rozen, S.G., Alexandrov, L.B.: Uncovering novel mutational signatures by de novo extraction with sigprofilerextractor. Cell Genomics **2**(11), 100179 (2022) https://doi.org/10.1016/j.xgen.2022.100179

[11] Jin, H., Gulhan, D.C., Ben-Isvy, D., Geng, D., Ljungstrom, V., Park, P.J.: Accurate and sensitive mutational signature analysis with musical. bioRxiv (2022) https://doi.org/10.1101/2022.04.21.489082 https://www.biorxiv.org/content/early/2022/04/22/2022.04.21.489082.full.pdf

[12] Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J., Stratton, M.R.: Deciphering signatures of mutational processes operative in human cancer. Cell Reports **3**(1), 246–259 (2013) https://doi.org/10.1016/j.celrep.2012.12.008

[13] Eigen, D., Ranzato, M., Sutskever, I.: Learning Factored Representations in a Deep Mixture of Experts (2014). https://arxiv.org/abs/1312.4314

[14] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NIPS-W (2017)

[15] Loshchilov, I., Hutter, F.: SGDR: Stochastic Gradient Descent with Warm Restarts (2017)

[16] Díaz-Gay, M., Vangara, R., Barnes, M., Wang, X., Islam, S.M.A., Vermes, I., Narasimman, N.B., Yang, T., Jiang, Z., Moody, S., Senkin, S., Brennan, P., Stratton, M.R., Alexandrov, L.B.: Assigning mutational signatures to individual samples and individual somatic mutations with sigprofilerassignment. bioRxiv (2023) https://doi.org/10.1101/2023.07.10.548264 https://www.biorxiv.org/content/early/2023/07/11/2023.07.10.548264.full.pdf

[17] Staaf, J., Glodzik, D., Bosch, A., Vallon-Christersson, J., Reuterswärd, C., Häkkinen, J., Degasperi, A., Amarante, T.D., Saal, L.H., Hegardt, C., Stobart, H., Ehinger, A., Larsson, C., Ryden, L., Loman, N., Malmberg, M., Kvist, A., Ehrencrona, H., Davies, H.R., Borg, Å., Nik-Zainal, S.: Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. Nature Medicine **25**(10), 1526–1533 (2019) https://doi.org/10.1038/s41591-019-0582-4

[18] Davies, H., Glodzik, D., Morganella, S., Yates, L.R., Staaf, J., Zou, X., Ramakrishna, M., Martin, S., Boyault, S., Sieuwerts, A.M., Simpson, P.T., King, T.A., Raine, K., Eyfjord, J.E., Kong, G., Borg, Å., Birney, E., Stunnenberg, H.G., Vijver, M.J., Børresen-Dale, A.-L., Martens, J.W.M., Span, P.N., Lakhani, S.R., Vincent-Salomon, A., Sotiriou, C., Tutt, A., Thompson, A.M., Van Laere, S., Richardson, A.L., Viari, A., Campbell, P.J., Stratton, M.R., Nik-Zainal, S.: Hrdetect is a predictor of brca1 and brca2 deficiency based on mutational signatures. Nature Medicine **23**(4), 517–525 (2017) https://doi.org/10.1038/nm.4292

[19] Cheng, D.T., Mitchell, T.N., Zehir, A., Shah, R.H., Benayed, R., Syed, A., Chandramohan, R., Liu, Z.Y., Won, H.H., Scott, S.N., Brannon, A.R., O'Reilly, C., Sadowska, J., Casanova, J., Yannes, A., Hechtman, J.F., Yao, J., Song, W., Ross, D.S., Oultache, A., Dogan, S., Borsu, L., Hameed, M., Nafa, K., Arcila, M.E., Ladanyi, M., Berger, M.F.: 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 (2015) https://doi.org/10.1016/j.jmoldx.2014.12.006

[20] Zehir, A., Benayed, R., Shah, R.H., Syed, A., Middha, S., Kim, H.R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S.M., Hellmann, M.D., Barron, D.A., Schram, A.M., Hameed, M., Dogan, S., Ross, D.S., Hechtman, J.F., DeLair, D.F., Yao, J., Mandelker, D.L., Cheng, D.T., Chandramohan, R., Mohanty, A.S., Ptashkin, R.N., Jayakumaran, G., Prasad, M., Syed, M.H., Rema, A.B., Liu, Z.Y., Nafa, K., Borsu, L., Sadowska, J., Casanova, J., Bacares, R., Kiecka, I.J., Razumova, A., Son, J.B., Stewart, L., Baldi, T., Mullaney, K.A., Al-Ahmadie, H., Vakiani, E., Abeshouse, A.A., Penson, A.V., Jonsson, P., Camacho, N., Chang, M.T., Won, H.H., Gross, B.E., Kundra, R., Heins, Z.J., Chen, H.-W., Phillips, S., Zhang, H., Wang, J., Ochoa, A., Wills, J., Eubank, M., Thomas, S.B., Gardos, S.M., Reales, D.N., Galle, J., Durany, R., Cambria, R., Abida, W., Cercek, A., Feldman, D.R., Gounder, M.M., Hakimi, A.A., Harding, J.J., Iyer, G., Janjigian, Y.Y., Jordan, E.J., Kelly, C.M., Lowery, M.A., Morris, L.G.T., Omuro, A.M., Raj, N., Razavi, P., Shoushtari, A.N., Shukla, N., Soumerai, T.E., Varghese, A.M., Yaeger, R., Coleman, J., Bochner, B., Riely, G.J., Saltz, L.B., Scher, H.I., Sabbatini, P.J., Robson, M.E., Klimstra, D.S., Taylor, B.S., Baselga, J., Schultz, N., Hyman, D.M., Arcila, M.E., Solit, D.B., Ladanyi, M., Berger, M.F.: Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nature Medicine **23**(6), 703–713 (2017) https://doi.org/10.1038/nm.4333

[21] Nguyen, B., Fong, C., Luthra, A., Smith, S.A., DiNatale, R.G., Nandakumar, S., Walch, H., Chatila, W.K., Madupuri, R., Kundra, R., Bielski, C.M., Mastrogiacomo, B., Donoghue, M.T.A., Boire, A., Chandarlapaty, S., Ganesh, K., Harding, J.J., Iacobuzio-Donahue, C.A., Razavi, P., Reznik, E., Rudin, C.M., Zamarin, D., Abida, W., Abou-Alfa, G.K., Aghajanian, C., Cercek, A., Chi, P., Feldman, D., Ho, A.L., Iyer, G., Janjigian, Y.Y., Morris, M., Motzer, R.J., O'Reilly, E.M., Postow, M.A., Raj, N.P., Riely, G.J., Robson, M.E., Rosenberg, J.E., Safonov, A., Shoushtari, A.N., Tap, W., Teo, M.Y., Varghese, A.M., Voss, M., Yaeger, R., Zauderer, M.G., Abu-Rustum, N., Garcia-Aguilar, J., Bochner, B., Hakimi, A., Jarnagin, W.R., Jones, D.R., Molena, D., Morris, L., Rios-Doria, E., Russo, P., Singer, S., Strong, V.E., Chakravarty, D., Ellenson, L.H., Gopalan, A., Reis-Filho, J.S., Weigelt, B., Ladanyi, M., Gonen, M., Shah, S.P., Massague, J., Gao, J., Zehir, A., Berger, M.F., Solit, D.B., Bakhoum, S.F., Sanchez-Vega, F., Schultz, N.: Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. Cell **185**(3), 563–57511

(2022) https://doi.org/10.1016/j.cell.2022.01.003

[22] Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A.P., Caldas, C., Davies, H.R., Desmedt, C., Eils, R., Eyfjörd, J.E., Foekens, J.A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinski, M., Jäger, N., Jones, D.T.W., Jones, D., Knappskog, S., Kool, M., Lakhani, S.R., López-Otín, C., Martin, S., Munshi, N.C., Nakamura, H., Northcott, P.A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J.V., Puente, X.S., Raine, K., Ramakrishna, M., Richardson, A.L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T.N., Span, P.N., Teague, J.W., Totoki, Y., Tutt, A.N.J., Valdés-Mas, R., Buuren, M.M., Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L.R., Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, PedBrain, I., Zucman-Rossi, J., Futreal, P.A., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S.M., Siebert, R., Campo, E., Shibata, T., Pfister, S.M., Campbell, P.J., Stratton, M.R.: Signatures of mutational processes in human cancer. Nature **500**(7463), 415–421 (2013)

[23] Alexandrov, L.B., Ju, Y.S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., Campbell, P.J., Vineis, P., Phillips, D.H., Stratton, M.R.: Mutational signatures associated with tobacco smoking in human cancer. Science **354**(6312), 618–622 (2016)