SigLASSO: a LASSO approach jointly optimizing sampling likelihood and cancer mutation signatures

Shantao Li[1,2], Forrest W. Crawford[3,4,5] & Mark B. Gerstein[1,2,6]

[1]*Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA.*

[2]*Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA.*

[3]*Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA.*

[4]*Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA.*

[5] *Yale School of Management, New Haven, Connecticut, USA.*

[6]*Department of Computer Science, Yale University, New Haven, Connecticut, USA.*

Multiple mutational processes drive carcinogenesis, leaving characteristic signatures on tumor genomes. Determining the active signatures from the full repertoire of potential ones can help elucidate the mechanisms underlying cancer initiation and development. This involves decomposing the frequency of cancer mutations categorized according to their trinucleotide context into a linear combination of known mutational signatures. We formulate this task as an optimization problem with L1 regularization and develop a software tool, sigLASSO, to carry it out efficiently. First, by explicitly adding multinomial sampling into the overall objective function, we jointly optimize the likelihood of sampling and signature fitting. This is especially important when mutation counts are low and sampling variance, high, such as the case in whole exome sequencing. sigLASSO uses L1 regularization to parsimoniously assign signatures to mutation profiles, leading to sparse and more biologically interpretable solu-

tions. Additionally, instead of hard thresholding and choosing a priori, a discrete subset of active signatures, sigLASSO fine-tunes model complexity parameters, informed by the scale of the data and prior knowledge. Finally, it is challenging to evaluate sigLASSO signature assignments. To do this, we construct a set of criteria, which we can apply consistently across assignments.

1 Main

Mutagenesis is a fundamental process underlying cancer development. Examples of mutational mechanisms include spontaneous deamination of cytosines, the formation of pyrimidine dimers by ultraviolet (UV) light, and the crosslinking of guanines by alkylating agents. Multiple endogenous and exogenous mutational processes drive cancer mutagenesis and leave distinct fingerprints (1). Notably, these processes have characteristic mutational nucleotide context biases (2; 3; 4; 5; 6). Mutation profiling of cancer samples at presentation has revealed that mutations accumulate over a lifetime; this includes somatic alterations generated by multiple mutational processes both before cancer initiation and during cancer development. In a generative model, multiple latent mutational processes generate mutations over time, drawing from their corresponding nucleotide context distributions ("mutation signature") (4; 5). Here a mutation signature is a multinomial probability distribution of mutations of various nucleotide contexts. In cancer samples, mutations from various mutational processes are mixed and observable by sequencing.

By applying unsupervised methods such as non-negative matrix factorization (NMF) and clustering to large-scale cancer studies, researchers have decomposed the mutation mixture and identified

at least 30 distinct mutational signatures (2; 7). Many signatures have been linked with mutational processes with known etiologies, such as aging, smoking, or ApoBEC activity. Investigating the fundamental processes underlying mutagenesis can help elucidate cancer initiation and development.

One major task in cancer research is to leverage signature studies on large-scale cancer cohorts and efficiently attribute active signatures to new cancer samples. A popular previously published method, deconstructSigs (8), decomposes the mutation profile into a signature mixture using binary search to try coefficients one-by-one iteratively and then by pruning signatures with low estimated contribution to archive sparsity. Other approaches use linear programming (9) or iterate all combinations by brute force (10). None of these approaches explicitly formulates sampling uncertainty into the model or uses efficient regression techniques.

Although we do not fully know the latent mutational processes in cancer samples, we can make reasonable and logical assumptions about the optimal solutions of such studies. Here, we aimed to design a computational framework, sigLASSO, that could meet several criteria. First, the set of estimated mutational mechanisms should be small, as past studies indicate that not all signatures can be active in a single sample or even a given cancer type. We also prefer a sparser solution as it explains an observation in a simpler fashion, consistent with Occam's razor principle. Second, the estimated mutational mechanisms should be biologically interpretable and reflect some cancer type specificity. For example, we should not observe UV-associated signatures in tissues that are not exposed to UV. Likewise, we only expect to observe activation-induced cytidine deaminase (AID) mutational processes, which are biologically involved in antibody diversification, in B-cell

lymphomas. Finally, the solution should be robust to the exact datasets for signatures and the data should control the model complexity.

In particular, it is a major challenge to reliably recover the signature composition when mutation number is low (8). Low mutation count results in high sampling variance, leading to an unreliable estimation of the mutation context probability distribution, which is the target for signature fitting. A desirable signature identification tool should model the sampling process and take sampling variance into consideration.

In this work, we formulate the task as a joint optimization problem with L1 regularization. First, by jointly fitting signatures and the parameters of a multinomial sampling process, sigLASSO takes into account the sampling uncertainty. Cooperatively fitting a linear mixture and maximizing the sampling likelihood enables knowledge transfer and improves performance. Specifically, signature fitting imposes constraints on the previously unconstrained multinomial sampling probability distribution. Conversely, a better estimation of the multinomial sampling probability helps signature fitting. This property is especially critical in high sampling variance settings, for example, when we only observe low mutation counts in whole exome sequencing (WES). Second, sigLASSO penalizes the model complexity and achieves variable selection by regularization. The most straightforward way to do this would be to use the L0 norm (cardinality of active signatures), but this approach cannot be effectively optimized. Conversely, using the L2 norm flattened out at small values leads to many tiny, non-zero coefficients, which are hard to interpret biologically. sigLASSO uses the L1 norm, which promotes sparsity. The L1 norm is convex, and thus allows efficient optimization (11; 12). Additionally, this approach is able to harmoniously integrate prior

biological knowledge into the solution by fine-tuning penalties on the coefficients. Compared with the approach of subsetting signatures before fitting, our soft thresholding method is more flexible for noise and unidentified signatures. Finally, sigLASSO is aware of data complexity such as mutational number and patterns in the observation. Our method is automatically parameterized empirically on performance, allowing data complexity to inform model complexity. In this way, our approach also promotes result reproducibility and fair comparison of datasets.

In sum, sigLASSO exploits constraints in signature identification and provides a robust framework for scientists to achieve biologically sound solutions. sigLASSO also can empower researchers to use and integrate their biological knowledge and expertise into the model. Unveiling the underlying mutational processes in cancer samples will enable us to recognize and quantify new mutagens, understand the mutagenesis and DNA repair processes, and develop new therapeutic strategies (9; 13; 14; 15; 16).

2 Results

The signature identification problem Mutational processes leave mutations in the genome within distinct nucleotide contexts. Specifically, we considered the mutant nucleotide context and looked one nucleotide ahead and behind. This divides mutations into 96 trinucleotide contexts. Each mutational process carries a unique signature, which is represented by a multinomial distribution over mutations of trinucleotide context (Fig. 1a).

Thirty signatures were identified by NMF (with Frobenius norm penalty) and clustering from large-scale pan-cancer analyses (2; 3). Here, our objective was to leverage the pan-cancer analysis and

decompose mutations from new samples into a linear combination of signatures. Mathematically, the problem is formulated as the following non-negative regression problem and maintains the original Frobenius norm:

$$\mathcal{W} = \text{argmin}_{\mathcal{W} \in R^+} ||\mathcal{M} - \mathcal{S}\mathcal{W}||_2^2$$

The mutation matrix, $\mathcal{M}$ contains mutations of each sample cataloged into 96 trinucleotide contexts. $m_i(i = 1 \ldots n) \in \mathcal{M}$ denotes the mutation count of the i^{th} category. $\mathcal{S}$ is a 96×30 signature matrix, containing the mutation probability in 96 trinucleotide contexts of the 30 signatures. $\mathcal{W}$ is the weights matrix, representing the contributions of 30 signatures in each sample.

Sampling variance In practice, this problem is optimized using continuous relaxation for efficiency and simplicity (8), ignoring the discrete nature of mutation counts. This approach essentially transforms observed mutations into a multinomial probability distribution, making model estimation insensitive to the total mutation count. Yet the total mutation count plays a critical role in inference. Assuming mutations are drawn from an underlying probability distribution, which is the mixture of several mutational signatures, the mutations follow a multinomial distribution. The total mutation count is the sample size of the distribution, thus affecting the variance of the inferred distribution.

For instance, 20 mutations within the 96 categories give us very little confidence in inferring the underlying mutation distribution. By constast, if we observed 2,000 mutations, we would have much higher confidence. Here, we aimed to use a likelihood-based approach to acknowledge the sampling variance and design a tool sensitive to the total mutation count.

sigLASSO model We divided the data generation process into two parts. First, multiple mutational signatures mix together to form an underlying mutation distribution. Second, we observed a set of categorical data (mutations), which is a realization of the underlying mutation distribution. We used $m_i (i = 1 \ ... \ n)$ to denote the mutation count of the ith category. $\overrightarrow{p}$ is the underlying mutation probability distribution with p_j denoting the probability of the j^{th} category.

To promote sparsity and interpretability of the solution, sigLASSO adds an L1 norm regularizer on the weights (i.e., coefficients) of the signatures. LASSO is mathematically justified and can be computationally solved efficiently (12). Adding an L1 norm regularizer is equivalent of placing a Laplacian prior on $\overrightarrow{w}$ (17). Thus, from this generative model, we can write down the likelihood function for one single sample.

$$\mathcal{L} = \mathcal{P}(\overrightarrow{m}|\mathcal{S}\overrightarrow{w}) = \mathcal{P}(\overrightarrow{m}|\overrightarrow{p})\mathcal{P}(\overrightarrow{p}|\mathcal{S}\overrightarrow{w})$$
$$\propto \frac{N!}{\prod_{i=1}^{n} m_i!} \prod_{i=1}^{n} p_i^{m_i} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\sum_{i=1}^{n}(p_i - \sum_{k=1}^{K} s_{ik} w_k)^2}{2\sigma^2}} \prod_{k=1}^{K} e^{-\lambda c_k w_k}$$

The log-likelihood function, which is our objective function to maximize, is given as following:

$$\ell \propto \sum_{i=1}^{n} \{m_i \log p_i - \frac{\alpha}{2}(p_i - \sum_{k=1}^{K} s_{ik} w_k)^2\} - \lambda \sum_{k=1}^{K} c_k w_k$$
$$s.t. \forall w_k \geq 0, \forall p_i \geq 0, \sum_{i=1}^{n} p_i = 1$$

Here, $\alpha = 1/\sigma^2$. We can infer α from the residual errors from linear regression. $\overrightarrow{c}$ is a vector of k penalty weights $(c_1, c_2, \ ... \ c_k)$, each indicating the strength to penalize the coefficient of a certain signature. This value should be tuned to reflect the level of confidence in prior knowledge. We

also used $\overrightarrow{c}$ to perform adaptive LASSO (18) by initializjng $\overrightarrow{c}$ to $1/\beta_{OLS}$, where β_{OLS} are the coefficients from nonnegative ordinary least square. Our aim was to obtain a less biased estimator by applying smaller penalties on larger values.

Optimizing sigLASSO The negative log likelihood is convex in respect to both $\overrightarrow{p}$ and $\overrightarrow{w}$ when evaluated individually. Hence the loss function is biconvex. Instead of using a generic optimizer, we exploited the biconvex nature of this problem and effectively optimized the function by Alternative Convex Search (ACS), which iteratively updates these two variables (19).

To begin the iteration, we initialized $\overrightarrow{m}$ using maximum likelihood estimation (MLE) and started with the $\overrightarrow{w}-$step. The $\overrightarrow{w}-$step is a nonnegative linear LASSO regression that can be efficiently solved by glmnet (18). λ is parameterized empirically (See Methods).

Next, we used the LASSO error variance estimator to estimate α (18). We solved the $\overrightarrow{p}$ with a Lagrange multiplier to maintain the linear summation constraint $\sum_{i=1}^{n} p_i = 1$. The nonnegative constraint of p_i is satisfied by only retaining the nonnegative root of the solution (see Methods).

In $\overrightarrow{p}-$step, we tried to estimate $\overrightarrow{p}$ that optimizes the multinomial likelihood while constraining it to be not too far away from the fitted $\overrightarrow{p}$. If we only used the point MLE of $\overrightarrow{p}$ based on sampling and did not perform the $\overrightarrow{p}-$step, the model would assume the sampling is perfect and become insensitive to the total mutation counts. The trade-off in the $\overrightarrow{p}-$step between the multinomial likelihood and the L2 loss reflects the sampling error. The sampling size (sum of m_i), the goodness of the signature fit (as reflected in α), and the overall shapes of $\overrightarrow{p}$ all affect the tension between sampling and linear fitting.

Algorithm 1 LIBRA algorithm

initialization:

$p_i^0 \leftarrow p_i^{mle} = \frac{m_i}{\sum_{i=1}^{n} m_i}$

$t \leftarrow 0$

loop:

$\overrightarrow{w}^{t+1} \leftarrow \underset{\overrightarrow{w}=w_1,w_2,..w_i,...w_n}{\text{argmax}} \frac{\alpha}{2} \sum_{i=1}^{n} (p_i^t - \sum_{k=1}^{K} s_{ik} w_k)^2 - \lambda \sum_{k=1}^{K} c_k w_k$

($\overrightarrow{w}$-step)

$\overrightarrow{p}^{t+1} \leftarrow \underset{\overrightarrow{p}=p_1,p_2,..p_i,...p_n}{\text{argmax}} \sum_{i=1}^{n} \{m_i \log p_i - \frac{\alpha}{2} (p_i - \sum_{k=1}^{K} s_{ik} w_k^{t+1})^2\}$

($\overrightarrow{p}$-step)

if $||\overrightarrow{p}^{t+1} - \overrightarrow{p}^{t}||_2 < \epsilon$ **then**

    break

$t \leftarrow t + 1$

return $\overrightarrow{w}^{t+1}$

sigLASSO is aware of the sampling variance Jointly optimizing both sampling process and signature fitting, sigLASSO is aware of the sampling variance and infers an underlying mutational context distribution $\overrightarrow{p}$. The underlying latent distribution is optimized in respect to both sampling likelihood and the linear fitting of signatures (Fig. 1b). In low mutation counts, the uncertainty in sampling increases and thus the estimated underlying distribution goes closer to the least square estimate (Fig. 1c). In contrast, when the total mutation count is high, the estimate of the distribution is closer to the MLE of the multinomial sampling process.

We illustrated how the mutation count affects the estimation of $\overrightarrow{p}$ using a simulated dataset (five signatures, noise level: 0.1, see Methods). When the sample size was small ($\leq$ 100), high uncertainty in sampling pushed the inferred underlying mutational distribution $\overrightarrow{p}$ far from the MLE in exchange for better signature fitting. When the sample size increased, lower variance in sampling dragged $\overrightarrow{p}$ close to the sampling MLE and forced the signatures to fit with larger error.

Because linear fitting and sampling likelihood optimization mutually inform each other, concurrently learning an auxiliary sampling likelihood improves performance. We compared the accuracy of the estimation of $\overrightarrow{p}$ with and without this joint optimization (Fig. 1d). As expected, $\overrightarrow{p}$ estimation in low mutation count performs worse. sigLASSO is able to achieve a lower MSE in both estimating $\overrightarrow{p}$ and the underlying true signature mixture.

Performance on simulated datasets We first evaluated sigLASSO on simulated datasets. Both sigLASSO and deconstructSigs performed better with higher mutation number and lower noise (Fig. 2a, S1). A decrease in mutation number leads to an increase of uncertainty in sampling, which is mostly negligible in the high mutation scenarios. As expected, the MSE jumped to the 0.05-0.3 range regardless of the noise level when the mutation number was low. Similar pattern is observed in support recovery (i.e. precision/recall/accurary). Thus, the error is dominated by undersampling rather than embedded noise. Overall, the performances, measured by MSE, of the two tools were comparable. sigLASSO produces a sparser solution. It maintains a higher precision level when mutation number decreases and/or noise increases. The precision of sigLASSO is less affected by the number of signatures. When the signature number is low (sparse settings), sigLASSO shows a better accuracy.

Using known signature to tune the weights boosts performance (Fig. 2b, S2). As the fraction of true signatures given as prior knowledge increased, the performance improved. When more false signatures were mixed with true signatures given as prior knowledge, the performance slowly deteriorated. Stronger priors had bigger effects to the solution as expected.

Evaluating criteria for signature assignment We next moved from synthetic datasets to real cancer mutational profiles. Real cancer mutational profiles are likely noisier than our simulation and exhibit a highly non-random distribution of signatures.

One of the limitations of cancer signature research is that the ground truth of real samples typically cannot be obtained. Previous large-scale signature studies largely relied on mutagen exposure asso-

ciation from patient records and biochemistry knowledge on mutagenesis. Here, we illustrated the outputs of different models and compared the results with existing signature knowledge. Although no gold standard exists to evaluate the performance, we do have a few reasonable expectations about the solution:

Sparsity: One or more signature should be active in a given cancer sample and type. However, not all signatures should be active. Mutational processes are discrete in nature and tied with certain endogenous and environmental factors. An obvious example is that the UV signature should not exist in unexposed tissues. Previous signature studies suggest a sparse distribution of signatures among cancer samples and types. Existing signature identifying methods aim to implicitly achieve sparse solutions by dropping signatures with small coefficients or pre-selection of the signature set for fitting.

Cancer type-specific signatures: We expected to find divergent signature distributions in different cancer types. Various tissues are exposed to diverse mutagens and undergo mutagenesis in dissimilar fashions. Signature patterns should be able to distinguish between cancer types. It is unrealistic to have the same or similar distribution of signatures in all cancer types, as they have divergent endogenous biological features and environmental exposures.

Robustness: Solutions should be robust and reproducible. Signatures are not orthogonal, thus simple regression might lead to solutions that change erratically when a small perturbation is made in the observation. Moreover, the solution should reflect the level of ascertainment. Especially in WES, low mutation count is often a severe obstacle for assigning signatures due to undersampling. Care should be taken to avoid overfitting. In particular, under low mutation count,

not all of the operative signatures would be reliably discovered.

Biological interpretability: The solution should be biological interpretable. Because of the biological nature of co-linearity in the signatures, simple mathematical optimization might pick the wrong signature. Even LASSO does not provide a guarantee to pick the correct predictor. Researchers now solve this problem by simply removing the majority of predictors they believe to be inactive. SigLASSO allows users to supply domain knowledge to guide the variable selection in a soft thresholding manner.

These expectations are not quantitative, but they help direct us to recognize the most plausible solution as well as the less favorable ones.

WGS scenario using renal cancer datasets We benchmarked the two methods using 35 WGS papillary kidney cancer samples (20) . The median mutation count was 4,528 (range: 912-9,257). We found that without prior knowledge, both sigLASSO and deconstructSigs showed high contributions from signature 3 and 5 (Fig. 3a,b). deconstructSigs also assigned a high proportion for signature 8, 9 and 16 . Signatures 3, 8, 9 and 16 were not found to be active in pRCC in previous studies and currently no biological support rationalizes their existence in pRCC (2).

However, if we naively subset the signatures and took the ones that were found to be active in previous studies, the signature profile was completely dominated by signature 5, to which only less than 5% mutations were assigned with other signature. This finding suggests possible an overly simply, underfitted model.

When sigLASSO took into account the prior knowledge of active signatures, the proportion of

backbone signature 5 increased to about 80%, which is in line with previous reports. SigLASSO also assigned a small portion of mutations to signature 3 and 13.

WES scenario using esophageal carcinoma datasets We next aimed to evaluate the two methods on 182 WES esophageal carcinoma (ESCA) samples with more than 20 mutations. The median mutation count was 175.5 (range: 28-2,146), which is considerably lower than WGS but typical for WES. We did not use any prior knowledge because COSMIC does not have active signatures in esophageal cancers.

In sigLASSO, the L1 penalty strength is tuned based on model performance. In a low mutation count, ESCA WES dataset, the model variance is likely high, which pushes up the penalty. As expected, sigLASSO assigned a lower fraction of mutations having signatures than deconstructSigs (median: 0.47 and 0.95; interquartile range: 0.36-0.56, 0.91-0.98, respectively, Fig. S3). The leading signatures were 25, 3, 1, and 9 in both tools.

DeconstructSigs has been shown to be able to distinguish between different histological types of esophageal cancer (8). We demonstrated that sigLASSO generates a sparser but comparable result with wider signature 3 and 25 gaps between the subtypes (Fig. 3c). The adenocarcinoma subtypes tended to have higher fractions of signature 1 and 25, and lower fractions of signature 3 and 9.

Performance on 8,892 The Cancer Genome Atlas (TCGA) samples We ran sigLASSO with step-by-step set-ups and deconstructSigs on 8,892 TCGA tumors (33 cancer types, S5) with more

than 20 mutations (Fig. 4).

We noticed that simple nonnegative regression resulted in an overly dense matrix. Applying an L1 penalty made the solution remarkablely sparse. Then by incorporating the prior knowledge, the signature landscape further changed without significantly affecting on the assignment sparsity. The solution is in consistent with the priors given. In comparison, the solution of deconstructSigs was less sparse.

sigLASSO is computationally efficient sigLASSO iteratively solves two convex problems. The $\overrightarrow{w}-$step can be solved using a very efficient coordinate descent algorithm (glmnet) (12). The $\overrightarrow{p}-$step is solved by a set of quadric equations. We observed empirically that the solution quickly converges in a few iterations even with extremely low mutation numbers (about 10). DeconstructSigs uses a binary search to try and tune coefficients by looping through all signatures at each iteration.

By profiling sigLASSO and deconstructSigs (Fig. 5), we noticed that neither total mutation numbers nor signature numbers does remarkably affect the running time of sigLASSO. With a high mutation number, sigLASSO is roughly 3-4 times faster than deconstructSigs; with low mutation number (50 mutations), these two tools show a comparable computation time. Noticeably, despite using less time, sigLASSO employs empirically parameterization which actually solves the signature fitting problems hundreds of times in a typical run.

Discussion

Studies decomposing cancer mutations into a linear combination of signatures have provided invaluable insights into cancer biology (4; 5; 6; 21). Through inferring mutational signatures and latent mutational processes, researchers have gained a better understanding of one of the fundamental driving forces of cancer initiation and development: mutagenesis.

How to leverage results from large-scale signature studies and apply them to a small set of incoming samples is a very practical problem for many researchers. Although it might seem to be a simple linear system problem at first, the core challenge is how to prevent over- and underfitting on only one single sample, often with very few mutations (especially in WES) and promote sparsity. First, under the current generative model, cancer draws mutations from a multinomial distribution of all active cancer signatures and then further draws from the multinomial nucleotide context distribution given by the signature. Mutations are first divided into several signatures and then categorized further into 96 types based on the nucleotide composition. With the mutation number less than a few hundred, sampling variance becomes a significant factor in reliable signature identification. Therefore, the fitting scheme should be aware of the sampling variance, which is especially pronounced in low mutation count scenarios (WES or cancer types with low mutation burden). Ideally, the tool should be able to attribute the signatures by flexibly inferring the underlying true mutation distribution given the sampling variance and the signature fitting performance. Second, the solution should be sparse. Signature studies on large-scale cancer datasets have revealed that mutational signatures are not all active in one sample or cancer type. In most tumor cases, only a few signatures prevail. A recent signature summary suggested that 2 to 13 known signatures

are observed in a given cancer type (based on 30 COMIC signatures), which might include hundreds and even thousands of samples. Sparse solutions are biologically sound and interpretable. In addition, sparse solutions can give better predictions due to lower estimator variance and are in line with Occam's razor principle, which prefers the simplest solution that explains an observation. Third, a desirable method should be aware of data complexity and be parameterized accordingly to achieve optimum fitting. Finally, mutational signatures are not orthogonal due to their biological nature. Co-linearity of the signatures will lead to unstable fittings that change erratically with even a slight perturbation of the observation.

DeconstructSigs was the first tool to identify signatures even in a single tumor. This tool uses binary search to iteratively tune coefficients and archives sparsity by post-hoc pruning with a preset 6% cut-off value. The mutation spectrum is normalized before fitting, thus making mutation counts irrelevant to the model. Moreover, the greedy nature of stepwise coefficient tunning is prone to eliminating valuable predictors in later steps that are correlated with previously selected ones (22). Here, we describe sigLASSO, which simultaneously optimizes both the sampling process and an L1 regularized signature fitting. By explicitly formulating a multinomial sampling likelihood into the optimization, we designed sigLASSO to take into account the sampling variance. Meanwhile, sigLASSO uses the L1 norm to penalize the coefficients, thus promoting sparsity and achieving effective variable selection. By fine-tuning the penalizing terms using prior biological knowledge, sigLASSO is able to further exploit previous signature studies from large cohorts and promote signatures that are believed to be active in a soft thresholding manner.

Jointly optimizing a mutation sampling process enables sigLASSO to be aware of the sampling

variance. By additionally modeling an auxiliary multinomial sampling process and corresponding distribution, we demonstrated that sigLASSO achieves better signature attribution, especially in low mutation counts cases. In cancer research, WES data is abundant but it also suffers severely from undersampling in signatures attribution. In these cases, sigLASSO is able to simultaneously learn the linear regression of signatures with a multinomial sampling process, generating more reliable and robust solutions. Moreover, formulating the mutation sampling uncertainty can have further implications in mutagenesis modeling and parameter estimation, for example in estimating the nucleotide-specific background mutation rate in cancer.

Additionally, as the cost of sequencing drops rapidly, we expect an even greater number of cancer samples to be whole-genome sequenced (23). The vast amount of cancer genomics data will give scientists larger power to discern unknown or rare signatures. The growing number of signatures will eventually make the signature matrix underdetermined (when $k > 96$, i.e., the number of possible mutational trinucleotide contexts). A traditional simple solver method would give infinitude (noiseless) or unstable (noisy) solutions in this underdetermined linear system. However, by assuming the solution is sparse, we were able to apply regulation to achieve a simpler, sparser and stable solution.

Moreover, sigLASSO does not specify a noise level explicitly beforehand, but instead empirically tunes parameters based on model performance. On the other hand, deconstructSigs specifies a noise level of 0.05 to derive the cut-off of 0.06 to achieve sparsity. In general, sigLASSO lets the data itself control the model complexity and leave any post-hoc filtering to users.

Finally, due to the colinearity nature of signatures, pure mathematical optimization might lead al-

gorithms to select wrong signatures that are highly correlated with truly active ones. To overcome this problem, sigLASSO allows researchers to incorporate domain knowledge to guide signature identification. This knowledge input could be cancer-type specific signatures or patient clinical information (e.g., smoking history or chemotherapy). We showcased the performance of sigLASSO on real cancer datasets. Although we lack the ground truth of the operative mutational signatures in tumors, we have several reasonable beliefs about the signature solution. sigLASSO produced signature solutions that are biologically interpretable, properly align with our current knowledge about mutational signatures, and well distinguish cancer types and histological subtypes.

Methods

Optimization of the $p-$step. In the $\overrightarrow{p}$-step, we tried to solve the following problem with $\tilde{p_i}$ from estimation of the $w-$step: $\tilde{p}_i^{t+1} = \sum_{k=1}^{K} s_{ik} w_k^{t+1}$.

$$\overrightarrow{p} = \underset{\overrightarrow{p}=p_1,p_2,..p_i,...p_n}{\operatorname{argmax}} \sum_{i=1}^{n} \{m_i \log p_i - \frac{\alpha}{2}(p_i - \tilde{p_i})^2\}$$

$$s.t. \forall p_i \geq 0, \sum_{i=1}^{n} p_i = 1$$

$$\tilde{p_i} = \sum_{k=1}^{K} s_{ik} w_k$$

We added the Lagrangian multiplier and take the derivatives in respect to p_i, $(i = 1, 2...n)$ and λ. Now we get $n + 1$ equations.

$$p_1^2 + (\frac{\lambda}{\alpha} - \tilde{p_1})p_1 - \frac{m_1}{\alpha} = 0$$

...

$$p_i^2 + (\frac{\lambda}{\alpha} - \tilde{p_i})p_i - \frac{m_i}{\alpha} = 0$$

...

$$p_n^2 + (\frac{\lambda}{\alpha} - \tilde{p_n})p_n - \frac{m_n}{\alpha} = 0$$

$$\sum_{i=1}^{n} p_i = 1$$

The roots of the quadratic equation are given by

$$p_i = \frac{(\tilde{p_i} - \frac{\lambda}{\alpha}) \pm \sqrt{(\tilde{p_i} - \frac{\lambda}{\alpha})^2 + 4\frac{m_i}{\alpha}}}{2}$$

Both $\alpha = 1/\sigma^2$ is strictly positive and m_i is nonnegative. Therefore, if $m_i = 0$, there exists only one zero root and $p_i = 0$ iff. $m_i = 0$. If $m_i > 0$, there are exactly one negative and one positive root. Since we require $\forall p_i \geq 0$, we only keep the positive root. The second derivative of the log-likelihood is $-\frac{m_i}{p_i} - \alpha$, which is strictly negative. Therefore, the root we found is a maximum.

We plugged all the roots in the last equation (ie. the linear constrain) and use the R function uniroot() to solve λ.

Parameter tuning. We tuned lambda by repeatedly splitting the nucleotide contexts into training and testing sets and testing the performance. Because mutations of the same single nucleotide substitution context are correlated, we split the data set into eight subsets. Each subset contained two of every single nucleotide substitutions. We then held one subset as the testing dataset and only fit the signatures on the remaining ones. After circling all eight subsets and repeating the process 20 times, we used the largest λ (which leads to a sparser solution) that gave MSE 0.5 SD from the minimum MSE. λ was tuned when $\overrightarrow{p}$ deviated far from the estimation from the previous round. By adaptively learning $\overrightarrow{p}$, sigLASSO avoids overestimating the errors in the signature fitting and thus allows a higher fraction of mutations to be assigned with signatures. We fixed λ when the deviation was small to avoid the inherited randomness in subsetting that affects convergence.

$\alpha = 1/\sigma^2$. σ^2 is estimated using $\sigma^2 = \frac{n}{(n-k)(MSE)}$, where k is the number of nonzero coefficients in the LASSO estimator (24). sigLASSO updates α after every LASSO linear fitting. To avoid grossly overestimating σ^2 (thus underestimating α) in the initial steps when p is far from the optimum, we set a minimum α value. By default min $\alpha = 400$. This can be further tuned based on the strength of prior believe of noise level.

Data simulation and model evaluation. First, we downloaded 30 previously identified signatures (http://cancer.sanger.ac.uk/cosmic/signatures). We created a simulated dataset by randomly and uniformly drawing two to eight signatures and corresponding weights (minimum: 0.02). The addictive Gaussian noise was simulated at various levels with a positive normal distribution on 25% trinucleotide contexts. Then, we summed all the signatures and noise to form a mutation

distribution. We sample mutations from this distribution with different mutation counts.

We ran deconstructSigs according to the original publication (8) and sigLASSO without prior knowledge of the underlying signature. To evaluate the performances, we compared the inferred signature distribution with the simulated distribution and calculated MSE. We also measured the number of false positive and false negative signatures in the solution (support recovery).

Illustrating on real datasets. To assess the performance of our method on real-world cancer datasets, we used somatic mutations from various cancer types from TCGA. We downloaded MAF files from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). A detailed list of files used in this study can be found in S5.

We compared the signature composition results with a previous pan-cancer signature analysis (http://cancer.sanger.ac.uk/cosmic/signatures). We also extracted prior knowledge on active signatures in various cancer types from this source.

sigLASSO software suite. sigLASSO accepts processed mutational spectrums. We provided simple scripts to help parse mutational spectrums from VCF files. sigLASSO allows users to specify biological priors (i.e., signatures that should be active or inactive) and their weights. sigLASSO uses 30 COSMIC signatures by default. Users are also given the option to supply customized signature files. sigLASSO is computationally efficient; using default settings, the program can successfully decompose a WGS cancer sample in less than a few seconds on a regular laptop. For time profiling purpose, we ran sigLASSO and deconstructSigs on an Intel Xeon E5-2660 (2.60

GHz) CPU. We employed the R package "microbenchmark" to profile the function call siglasso() and whichSignatures(). For each setup, we generated ten noiseless simulated data sets and repeated 100 times for each evaluation.

We made sigLASSO code available on GitHub (https://github.com/gersteinlab/sigLASSO).

1. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).

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

3. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415 (2013).

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

5. Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Current opinion in genetics & development* **24**, 52–60 (2014).

6. Petljak, M. & Alexandrov, L. B. Understanding mutagenesis through delineation of mutational signatures in human cancer. *Carcinogenesis* **37**, 531–540 (2016).

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

8. 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**, 31 (2016).

9. Alexandrov, L. B., Nik-Zainal, S., Siu, H. C., Leung, S. Y. & Stratton, M. R. A mutational signature in gastric cancer suggests therapeutic strategies. *Nature communications* **6**, 8683 (2015).

10. Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nature genetics* **48**, 126 (2016).

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

12. Friedman, J., Hastie, T. & Tibshirani, R. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version* **1** (2009).

13. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618–622 (2016).

14. Viel, A. *et al.* A specific mutational signature associated with dna 8-oxoguanine persistence in mutyh-defective colorectal cancer. *EBioMedicine* **20**, 39–49 (2017).

15. Schulze, K. *et al.* Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nature genetics* **47**, 505 (2015).

16. Davies, H. *et al.* Hrdetect is a predictor of brca1 and brca2 deficiency based on mutational signatures. *Nature medicine* **23**, 517 (2017).

17. Park, T. & Casella, G. The bayesian lasso. *Journal of the American Statistical Association* **103**, 681–686 (2008).

18. Zou, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101**, 1418–1429 (2006).

19. Gorski, J., Pfeuffer, F. & Klamroth, K. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical methods of operations research* **66**, 373–407 (2007).

20. Li, S., Shuch, B. M. & Gerstein, M. B. Whole-genome analysis of papillary kidney cancer finds significant noncoding alterations. *PLoS genetics* **13**, e1006685 (2017).

21. Bailey, M. H. *et al.* Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385 (2018).

22. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. *et al.* Least angle regression. *The Annals of statistics* **32**, 407–499 (2004).

23. Muir, P. *et al.* The real cost of sequencing: scaling computation to keep pace with data generation. *Genome biology* **17**, 53 (2016).

24. Reid, S., Tibshirani, R. & Friedman, J. A study of error variance estimation in lasso regression. *Statistica Sinica* 35–67 (2016).

Acknowledgements We thank Dr. Harrison, Zhou, Dr. Cong Liang, Dr. Jinrui Xu, Diego Galeano, Mengting Gu for helpful discussion.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence should be addressed to M.B.G. (email: pi@gersteinlab.org).

Fig. 1: sigLASSO takes sampling variance into account

a. A schematic graph showing the mixture model of mutational processes and signatures.

b: A contour plot of the penalty function of multinomial sampling function (optimum at p_1) and the least square of signature fitting (optimum at p_2). sigLASSO tries to infer p by jointly optimizing both penalties (red contour lines, optimum at $\overrightarrow{p}$).

c: As mutation number increases, the inferred $\overrightarrow{p}$ gets closer to the sampling MLE rather than the linear fitting as the variance due to smaller sampling.

b: The MSE of the estimation of $\overrightarrow{p}$ and the underlying noiseless signature mixture by sigLASSO and using the point MLE. Low mutation counts profiles benefit from sigLASSO the most.

Fig 2: Performance on simulated datasets

a: The MSE of sigLASSO and using the point MLE to fit the signatures. Low mutation counts profiles benefit from sigLASSO the most.

b: MSE of the estimation of p by sigLASSO and using the point MLE. Again, low mutation counts profiles benefit from sigLASSO the most.

Fig 3: Performance on pRCC and ESCA samples

a. Signature assignment for 35 WGS pRCC samples. Bar plots show the fractions of mutation signature assignment for each sample using sigLASSO, sigLASSO without prior knowledge and deconstructSigs, and simple nonnegative regression.

b. A dot chart showing the mean fraction of mutation signatures in each sample. Signatures that contributed less than 0.05 are not shown.

c. A dot chart showing the mean fraction of mutation signatures in each sample, grouped by two tools and histological subtypes (adenocarcinoma/squamous). Signatures that contributed individually less than 0.05 in all four cases are grouped into "others".

Fig 4: Performance on 33 TCGA cancers

Active signatures in 33 cancer types using different methods. The numbers below cancer types are the counts of active signatures in each cancer type. Only 26 cancer types have previously known signature distributions (others are shaded). Signatures that contributes less than 0.5% are not shown.

Fig. 5: Running time of sigLASSO and deconstructSigs at different total mutations numbers

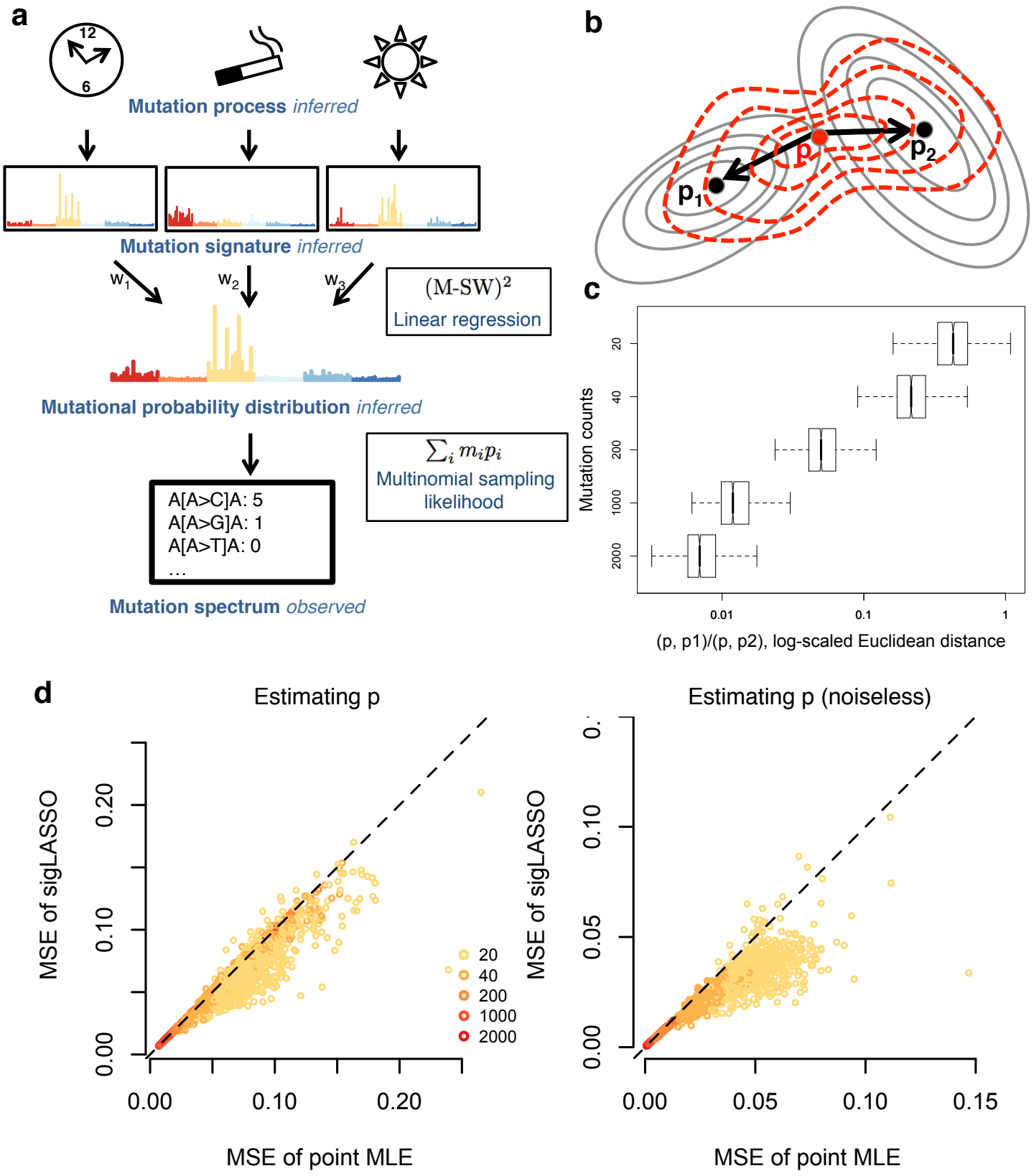
S1: Boxplots of MSE on simulated datasets. Red: sigLASSO, grey: deconstructSigs.

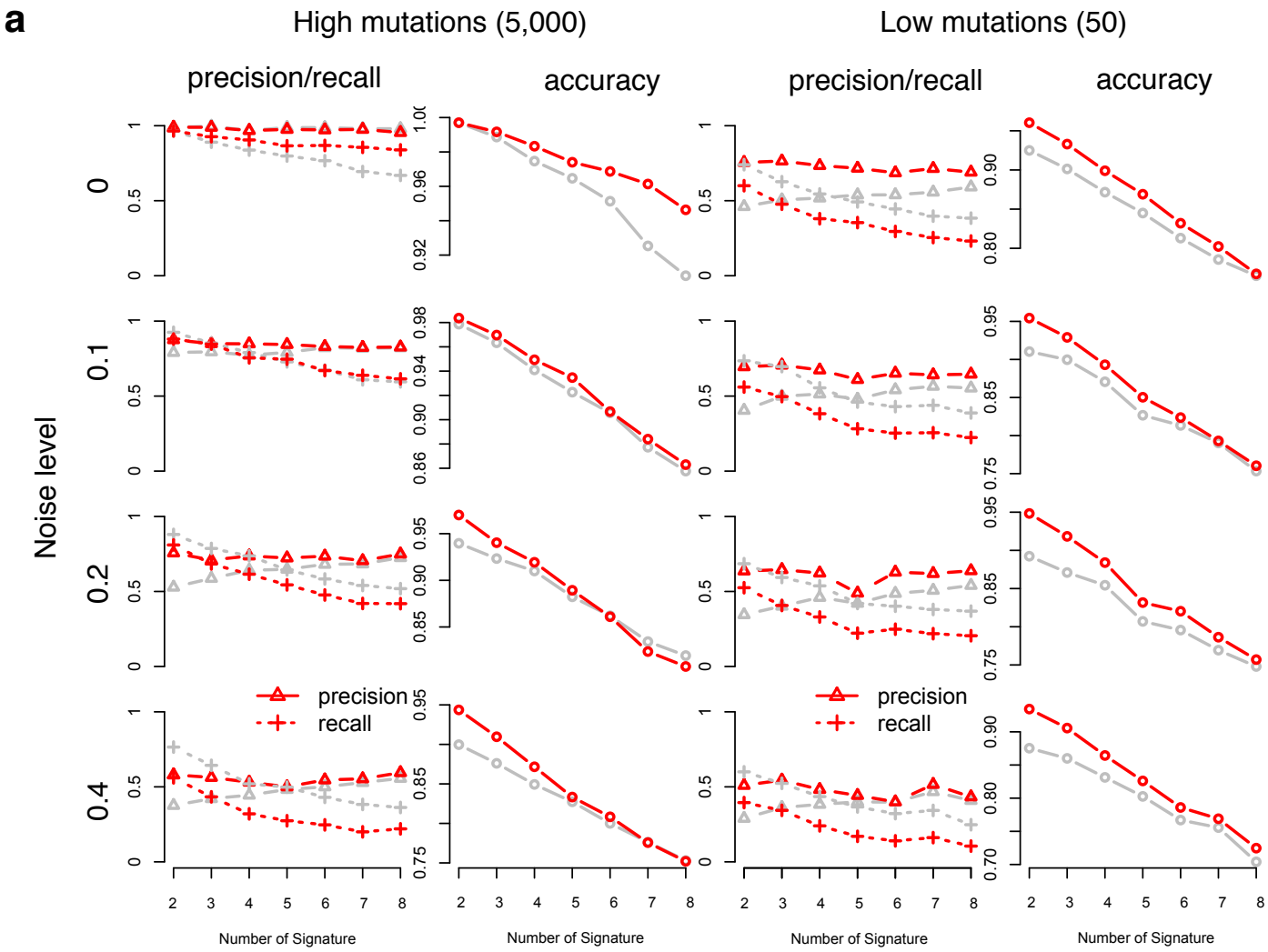
S2: MSE on simulated datasets, showing that tuning the penalty weights using prior knowledge improves performance. Penalty weights used: red, 0.5; yellow, 0.2; green, 0.1.

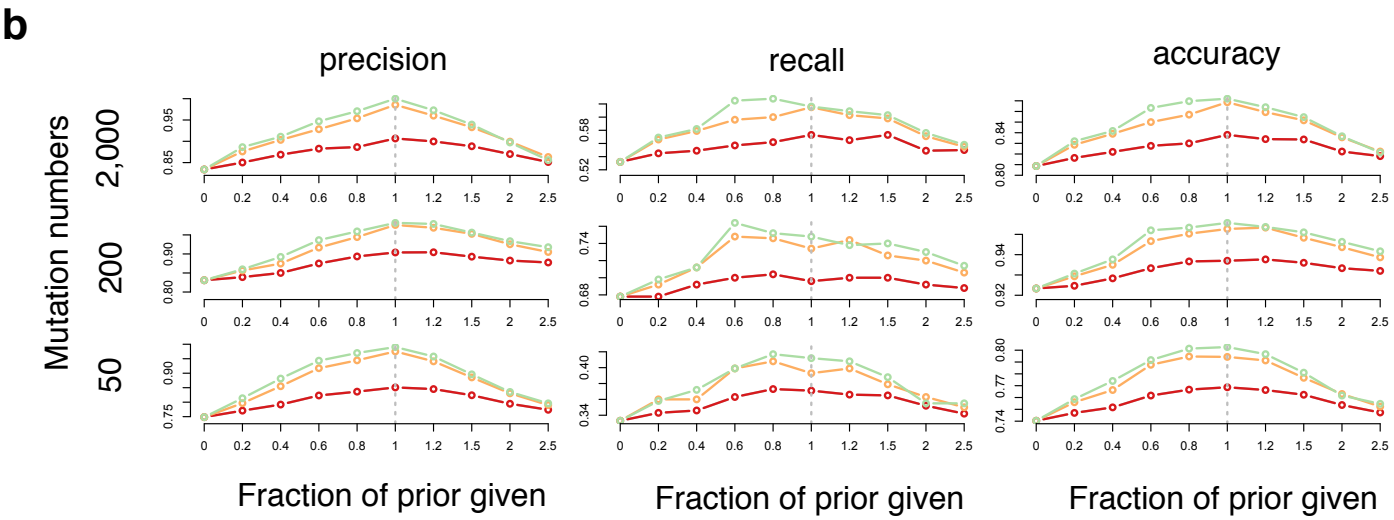
S3: Signature assignment for 182 WES ESCA samples. Bar plots show the fractions of mutation signature assignment for each sample using sigLASSO and deconstructSigs, sorted by the total mutation counts.

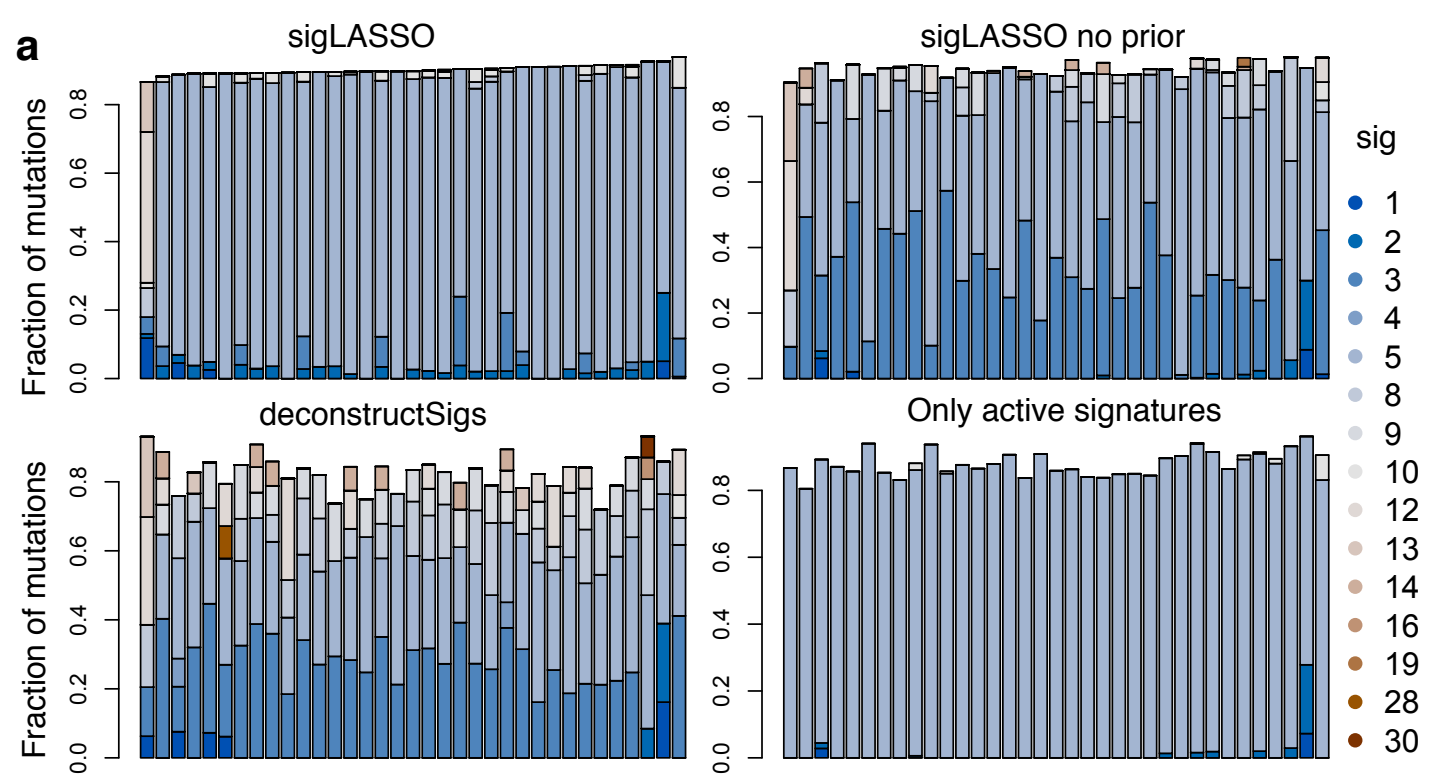
S4: Running time of sigLASSO at different numbers of signatures (downsampled).

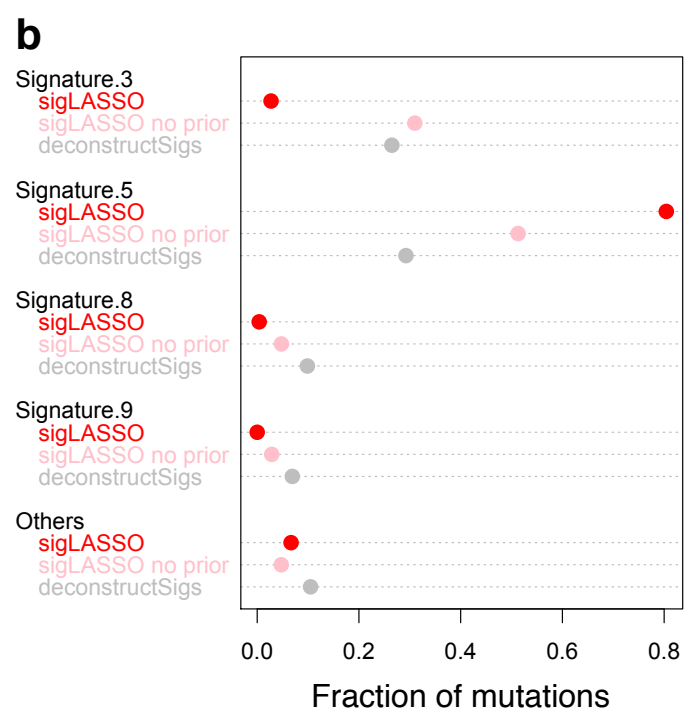
S5: A list of TCGA MAF files used in this study.

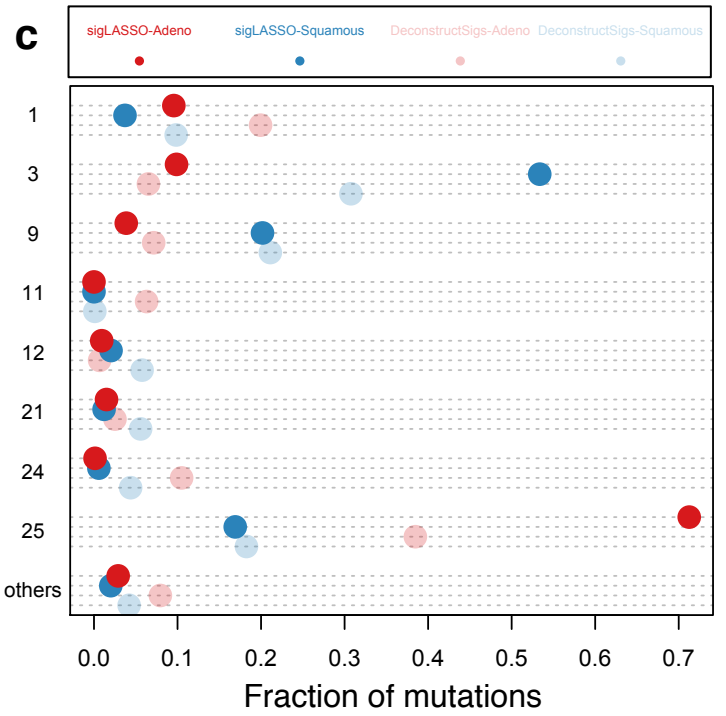

a
Mutation process inferred
Mutation signature inferred
w1
w2
w3
(M-SW)2
Linear regression
Mutational probability distribution inferred
∑i mipi
Multinomial sampling likelihood
A[A>C]A: 5
A[A>G]A: 1
A[A>T]A: 0
…
Mutation spectrum observed
b
p1
p
p2
c
Mutation counts
20
40
200
1000
2000
0.01
0.1
1
(p, p1)/(p, p2), log-scaled Euclidean distance
d
Estimating p
MSE of sigLASSO
MSE of point MLE
0.00
0.10
0.20
20
40
200
1000
2000
Estimating p (noiseless)
0.05
0.15

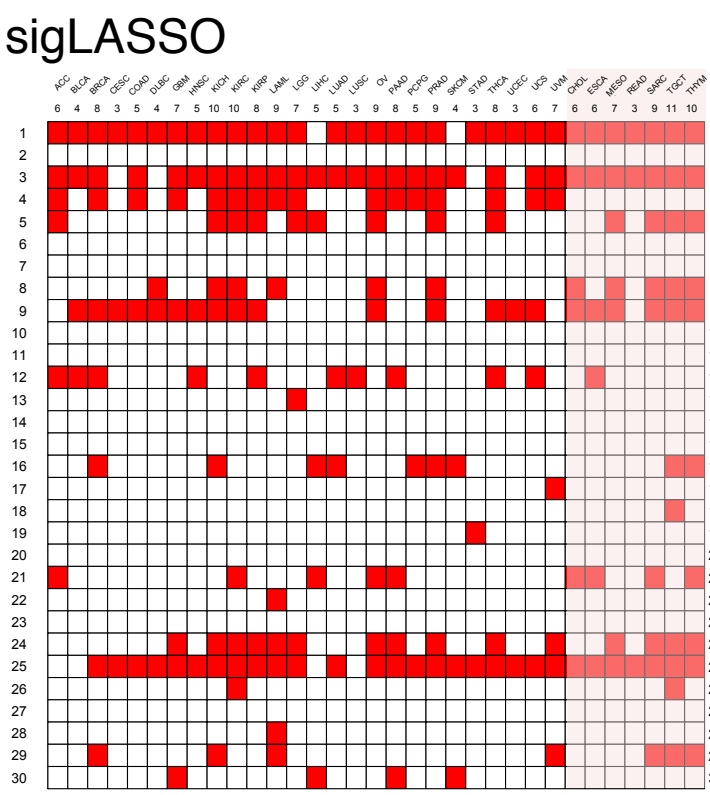

a
High mutations (5,000)
Low mutations (50)
precision/recall
accuracy
precision/recall
accuracy
Noise level
0
0.1
0.2
0.4
precision
recall
precision
recall
Number of Signature
Number of Signature
Number of Signature
Number of Signature

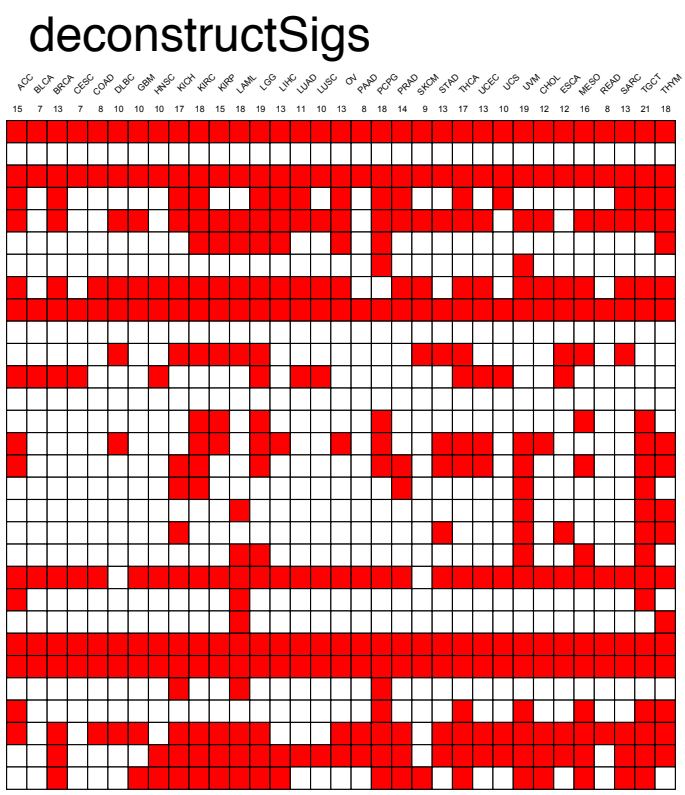

b
precision
recall
accuracy
Mutation numbers
2,000
200
50
Fraction of prior given
Fraction of prior given
Fraction of prior given


a
sigLASSO
sigLASSO no prior
deconstructSigs
Only active signatures
Fraction of mutations
0.0
0.2
0.4
0.6
0.8
sig
1
2
3
4
5
8
9
10
12
13
14
16
19
28
30


b
Signature.3
sigLASSO
sigLASSO no prior
deconstructSigs
Signature.5
sigLASSO
sigLASSO no prior
deconstructSigs
Signature.8
sigLASSO
sigLASSO no prior
deconstructSigs
Signature.9
sigLASSO
sigLASSO no prior
deconstructSigs
Others
sigLASSO
sigLASSO no prior
deconstructSigs
0.0
0.2
0.4
0.6
0.8
Fraction of mutations


c
sigLASSO-Adeno
sigLASSO-Squamous
DeconstructSigs-Adeno
DeconstructSigs-Squamous
1
3
9
11
12
21
24
25
others
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Fraction of mutations

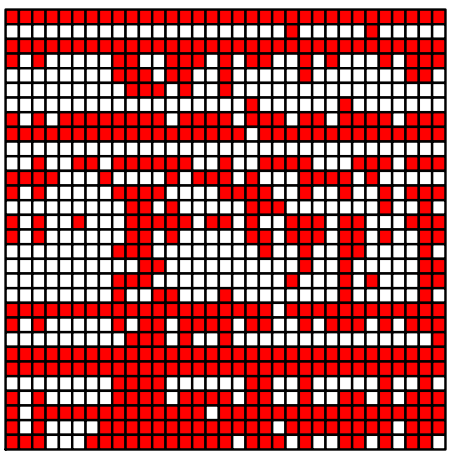


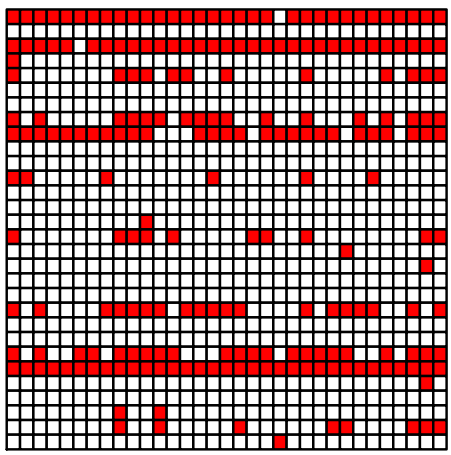


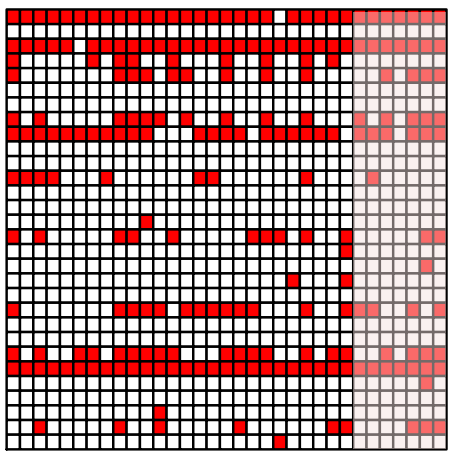
Simple regression

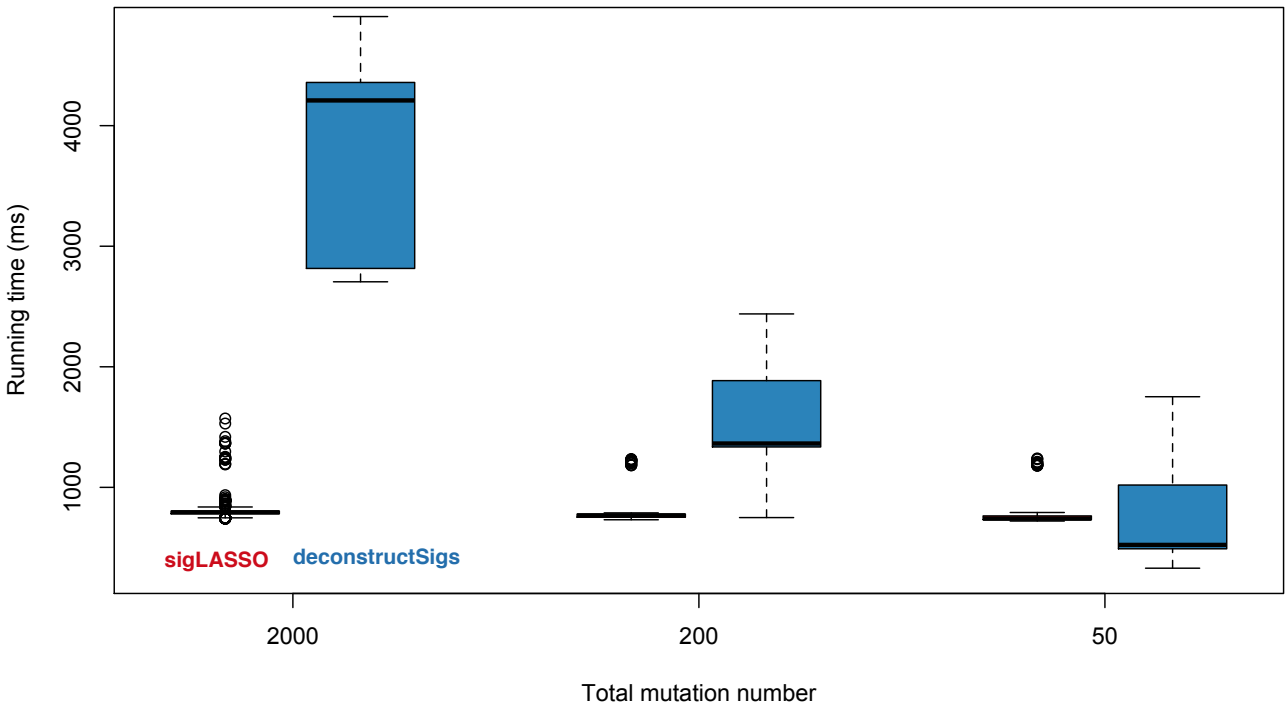
sigLASSO no prior

sigLASSO, weak prior








Running time (ms)
1000
2000
3000
4000
sigLASSO
deconstructSigs
2000
200
50
Total mutation number