nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-17388-x OPEN

Using sigLASSO to optimize cancer mutation signatures jointly with sampling likelihood

Shantao Li[1,2], Forrest W. Crawford[3,4,5,6] & Mark B. Gerstein[1,2,6,7]✉

Multiple mutational processes drive carcinogenesis, leaving characteristic signatures in tumor genomes. Determining the active signatures from a full repertoire of potential ones helps elucidate mechanisms of cancer development. This involves optimally decomposing the counts of cancer mutations, tabulated according to their trinucleotide context, into a linear combination of known signatures. Here, we develop sigLASSO (a software tool at github.com/gersteinlab/siglasso) to carry out this optimization efficiently. sigLASSO has four key aspects: (1) It jointly optimizes the likelihood of sampling and signature fitting, by explicitly factoring multinomial sampling into the objective function. This is particularly important when mutation counts are low and sampling variance is high (e.g., in exome sequencing). (2) sigLASSO uses L1 regularization to parsimoniously assign signatures, leading to sparse and interpretable solutions. (3) It fine-tunes model complexity, informed by data scale and biological priors. (4) Consequently, sigLASSO can assess model uncertainty and abstain from making assignments in low-confidence contexts.

[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, CT, USA. [6] Department of Statistics and Data Science, Yale University, New Haven, CT, USA. [7] Department of Computer Science, Yale University, New Haven, CT, USA. ✉email: mark@gersteinlab.org

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[2–6]. Sequencing cancer samples at presentation revealed all mutations accumulated over lifetime; these include 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 a set of 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 could help elucidate the initiation and development of cancer.

A major task in cancer research is to leverage signature studies on large-scale cancer cohorts and efficiently select and 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 iteratively test coefficients one-by-one and then hard pruning signatures with low estimated contribution to achieve 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. Moreover, no off-the-shelf implementation of these methods, besides deconstructSigs, is available.

Although we do not fully know the latent mutational processes in cancer samples, we can make reasonable and logical assumptions that facilitate our method design. Here, we aimed to design a computational framework, sigLASSO, which could meet these criteria. First, we assumed that the set of estimated mutational mechanisms should be small, as de novo studies indicate that not all signatures can be active in a single sample or even a given cancer type. In most cancer samples, only a few signatures are identified in the original de novo discovery studies. An ideal tool should generate solutions that follow this sparse distribution. Moreover, the number of detectable mutation signatures is limited by the amount of data support. Too many signatures lead to overfitting and unstable solutions. We aimed for a sparser solution as it explains observations in a simpler fashion. 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 mutational processes, which are biologically involved in antibody diversification, in B-cell lymphomas. Finally, we felt the solution should be robust and the data should control the model complexity.

In particular, reliably recovering the signature composition is challenging when the 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 formulated 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 improves 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). Existing methods use continuous relaxation, which makes the model invariant to different mutation counts. Second, sigLASSO penalizes the model complexity and achieves variable selection by regularization. Regularization is essential for this fitting problem with a large amount of signatures to achieve proper variable selection and avoid overfitting. Using regularization to promote sparsity and prevent overfitting is also a standard practice in de novo signature discovery. A few recent examples are rule-based constraints (SigProfiler), a Bayesian variant of NMF that penalizes on model complexity (SignatureAnalyzer[11]) and L1-based regularization (SparseSignatures[12]). 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 do not resemble the sparse signature distribution in the original de novo studies and are hard to interpret biologically. sigLASSO uses the L1 norm, which promotes sparsity. The L1 norm is convex, and thus allows efficient optimization[13,14]. In addition, 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 to noise and unidentified signatures. Finally, sigLASSO is aware of data complexity such as mutational number and patterns in the observation. It is able to abstain (decline to assign mutational processes under high uncertainty) and defer to the human researcher to decide. 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 data sets.

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 mutagenesis and DNA repair processes, and develop new therapeutic strategies for cancer[9,15–18].

Results

The signature identification problem. Mutational processes leave mutations in the genome within distinct nucleotide contexts. We denoted the total number of contexts as n. Typically, we considered the mutant nucleotide context and looked one nucleotide ahead and behind each mutation, dividing the mutations into $n = 96$ trinucleotide contexts. Each mutational process carries a unique signature, which is represented by a multinomial probability distribution over mutations of trinucleotide context (Fig 1a).

Then there are K latent mutational processes. Large-scale pan-cancer analyses identified $K = 30$ COSMIC signatures by NMF (with Frobenius norm penalty) and clustering[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, we formulated the following non-negative regression problem, maintaining the original Frobenius norm:

$$\mathcal{W} = \arg \min_{\mathcal{W} \in R^{+}} \|\mathcal{M} - \mathcal{S}\mathcal{W}\|_2^2 \tag{1}$$

The mutation matrix, $\mathcal{M}$ contains mutations of each sample cataloged into n trinucleotide contexts. $m_{i(i=1\ldots n)} \in \mathcal{M}$ denotes the mutation count of the ith category. $\mathcal{S}$ is a $n \times K$ 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] neglecting 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. However, the total mutation count has a critical role in inference. Assuming mutations are drawn from a latent 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 greatly 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 contrast, if we observed 2000 mutations, we would have much higher confidence. Methods using continuous relaxation treat these two conditions indifferently. 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 latent mutation distribution. Second, we observed a set of categorical data (mutations), which is a realization of the underlying mutation distribution. We used $m_i(i = 1\ldots n)$ to denote the mutation count of the ith category. The vector $\vec{p}$ is the underlying latent mutation probability distribution with p_j, denoting the probability of the jth category. The total number of mutations is N.

To achieve variable selection and promote sparsity and interpretability of the solution, sigLASSO adds an L1 norm regularizer on the weights $\vec{w}$ (i.e., coefficients) of the signatures with a hyperparameter λ. LASSO is mathematically justified and can be computationally solved efficiently[14]. Adding an L1 norm regularizer is equivalent to placing a Laplacian prior on $\vec{w}$[19]. $\vec{c}$ 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 vector should be tuned to reflect the level of confidence in the prior knowledge. For example, a smaller penalty weight represents a stronger prior, reflecting higher confidence and vice versa.

Overall, from the generative model, we can determine the likelihood function for a single sample.

$$\begin{aligned}\mathcal{L} &= \mathcal{P}(\vec{m}|\mathcal{S}\vec{w})\underbrace{\mathcal{P}(\vec{w})}_{\text{prior}}\\ &= \mathcal{P}(\vec{m}|\vec{p})\mathcal{P}(\vec{p}|\mathcal{S}\vec{w})\mathcal{P}(\vec{w})\\ &= \underbrace{\frac{N!}{\prod_{i=1}^{n} m_i!}\prod_{i=1}^{n} p_i^{m_i}}_{\text{multinomial sampling}} \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\sum_{i=1}^{n}\left(p_i - \sum_{k=1}^{K} s_{ik}w_k\right)^2}{2\sigma^2}}}_{\text{linear fitting}} \underbrace{\prod_{k=1}^{K} e^{-\lambda c_k w_k}}_{\text{Laplacian prior}}\end{aligned} \tag{2}$$

Our objective function is to maximize the log-likelihood function, which is given as:

$$\begin{gathered}\ell \propto \sum_{i=1}^{n}\{m_i \log \mathrm{p_i} - \tfrac{\alpha}{2}\left(\mathrm{p_i} - \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{s_{ik}w_k}\right)^2\} - \lambda \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{c_k w_k}\\ s.t. \forall w_k \geq 0, \forall p_i \geq 0, \sum_{i=1}^{n} p_i = 1\end{gathered} \tag{3}$$

Here, $\alpha = 1/\sigma^2$. We can infer α from the residual errors from linear regression (see parameter tuning). Meanwhile, because of its continuous nature, $\vec{c}$ can also be effectively learned using patient information (e.g., smoking status, tumor size, or methylation status). We also used $\vec{c}$ to perform adaptive LASSO[20] by initializing $\vec{c}$ to $1/\beta_{OLS}$, where β_{OLS} are the coefficients from non-negative ordinary least square. Our aim was to obtain a less-biased estimator by applying smaller penalties on variables with larger coefficients.

sigLASSO is aware of the sampling variance. By jointly optimizing both the sampling process and signature fitting, sigLASSO is aware of the sampling variance and infers an underlying mutational context distribution $\vec{p}$. The underlying latent distribution is optimized with 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 moves 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 $\vec{p}$ using a simulated data set (five signatures, noise level: 0.1, see "Methods"). When the sample size was small (≤100), high uncertainty in sampling pushed the inferred underlying mutational distribution $\vec{p}$ far from the MLE in exchange for better signature fitting. When the sample size increased, lower variance in sampling dragged $\vec{p}$ close to the sampling MLE and forced the signatures to fit even with larger errors.

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 $\vec{p}$ with and without this joint optimization (Fig. 1d). As expected, $\vec{p}$ estimation in low mutation count performed worse. sigLASSO was able to achieve a lower MSE in estimating both $\vec{p}$ (with noise) and the underlying true signature mixture (noiseless).

Performance on simulated data sets. We first evaluated sigLASSO on simulated data sets. Both sigLASSO (with and without priors) and deconstructSigs performed better with higher mutation number and lower noise (Fig. 2a, Supplementary Figure 1). A decrease in mutation number leads to an

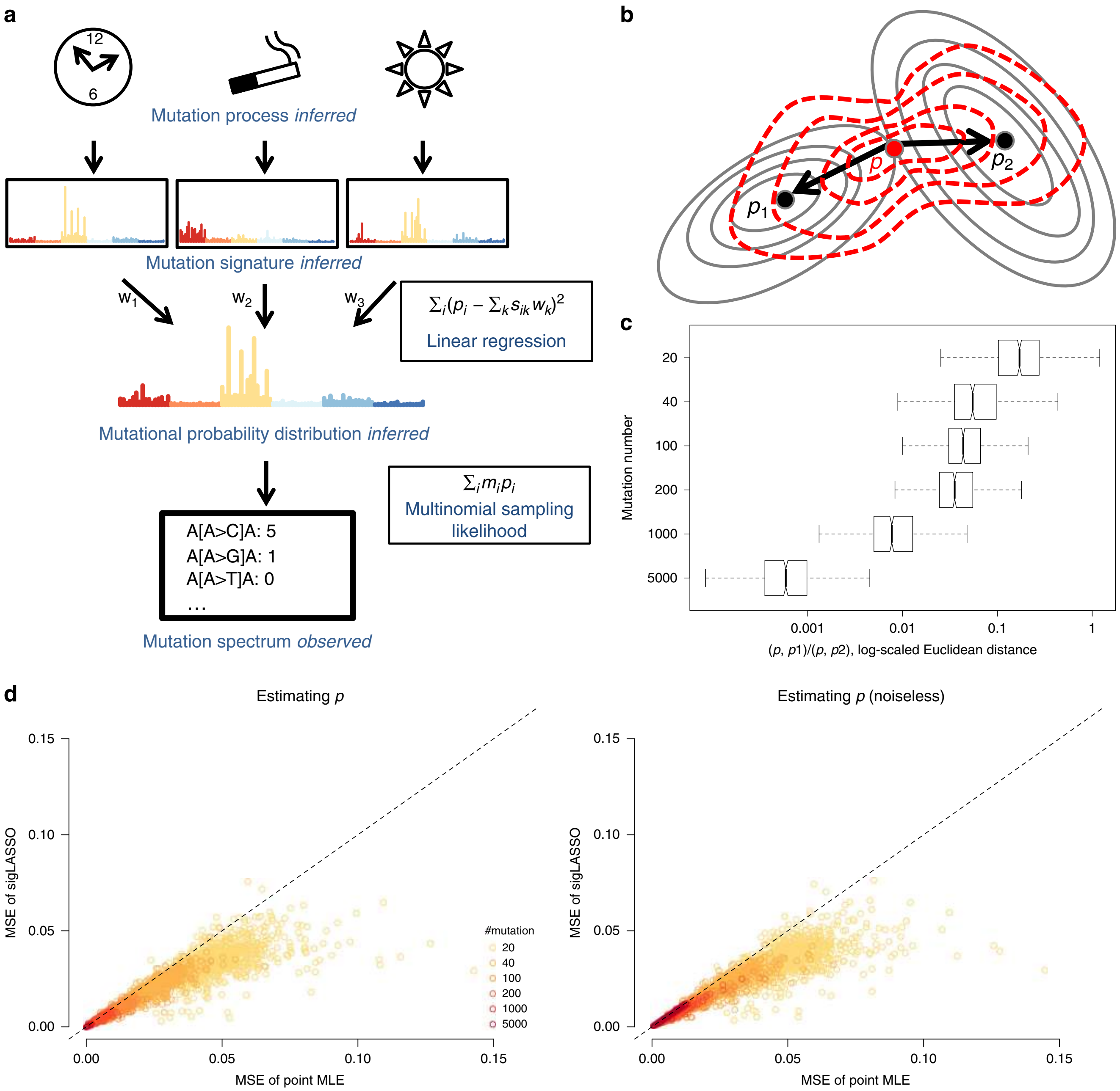

Fig. 1 SigLASSO takes sampling variance into account. a A schematic graph showing the mixture model of mutational processes and signatures. **b** A contour plot of the penalty function of multinomial sampling function (optimum at p_1) and the least square of signature fitting (optimum at p_2). sigLASSO tries to jointly optimize both penalties (red contour lines, optimum at p). **c** As mutation number increases, the inferred $\vec{p}$ gets closer to the sampling MLE rather than the linear fitting as the sampling variance decreases. Box edges are the 25th and 75th percentiles and whiskers indicate the 1.5× interquartile range (IQR) or max/min, which ones are smaller. **d** The MSE of the estimation of $\vec{p}$ and the underlying noiseless signature mixture by sigLASSO and using the point MLE. Low mutation counts profiles benefit from sigLASSO the most. Priors were sampled uniformly from the ground true positives and negatives.

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. We observed a similar pattern in support recovery (i.e., precision/recall/accuracy). Thus, the error is dominated by undersampling rather than embedded noise. Despite giving a simpler, sparse solution, sigLASSO with priors outperformed sigLASSO with no priors and deconstructSigs in both MSE and support recovery. In support recovery, sigLASSO with priors showed a 5–10% gain in accuracy compared with deconstructSigs. sigLASSO with no priors also had better support recovery performance, measured by accuracy, than deconstructSigs. Overall, sigLASSO maintained a higher precision level when the mutation number decreased and/or the noise increased, which shows its ability to abstain (decline to assign signatures) and provide some control on the false positive rate. Notably, sigLASSO with priors maintained an accuracy above 0.8 in all simulation settings. Moreover, the precision of sigLASSO was minimally affected by the number of signatures.

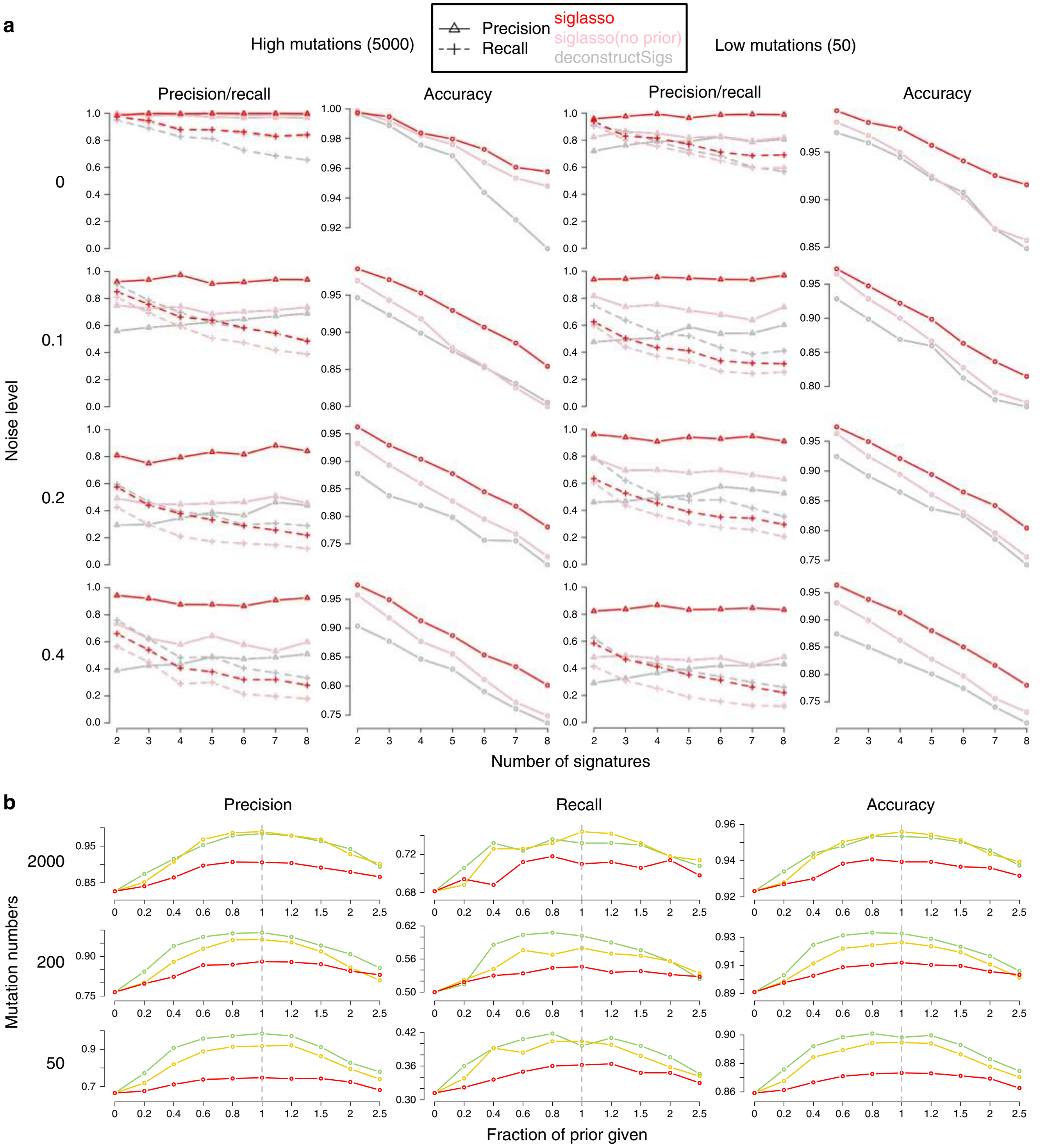

Fig. 2 Performance on simulated data sets. a Support recovery on simulated data sets. Each setting is simulated 100 times. Penalty coefficients for priors: 0.1. **b** Support recovery on simulated data sets. Tuning the penalty weights using prior knowledge improves performance; *x* axis: the fraction of true signatures given as prior (larger than 1 indicates false signatures giving as priors). Penalty coefficients for priors: red, 0.5; yellow, 0.1; green, 0.01.

Although the recall was slightly lower with sigLASSO than deconstructSigs in noisy settings, in very low and no noise situations sigLASSO achieved, a higher recall owing to its ability to adapt to different noise levels. By contrast, deconstructSigs, which assumes a fixed noise level, overly pruned the signatures in the post-fitting step when the noise was low. At last, adding correct priors helped boost both precision and recall significantly.

We next explored how different priors affect sigLASSO's performance. Our experiments showed that using known signatures as priors to tune the weights boosts performance. Priors improved performance even when we included only a small fraction of true signatures or blended in a large number of wrong signatures (Fig. 2b, Supplementary Figure 2). As the fraction of true signatures given as prior knowledge increased

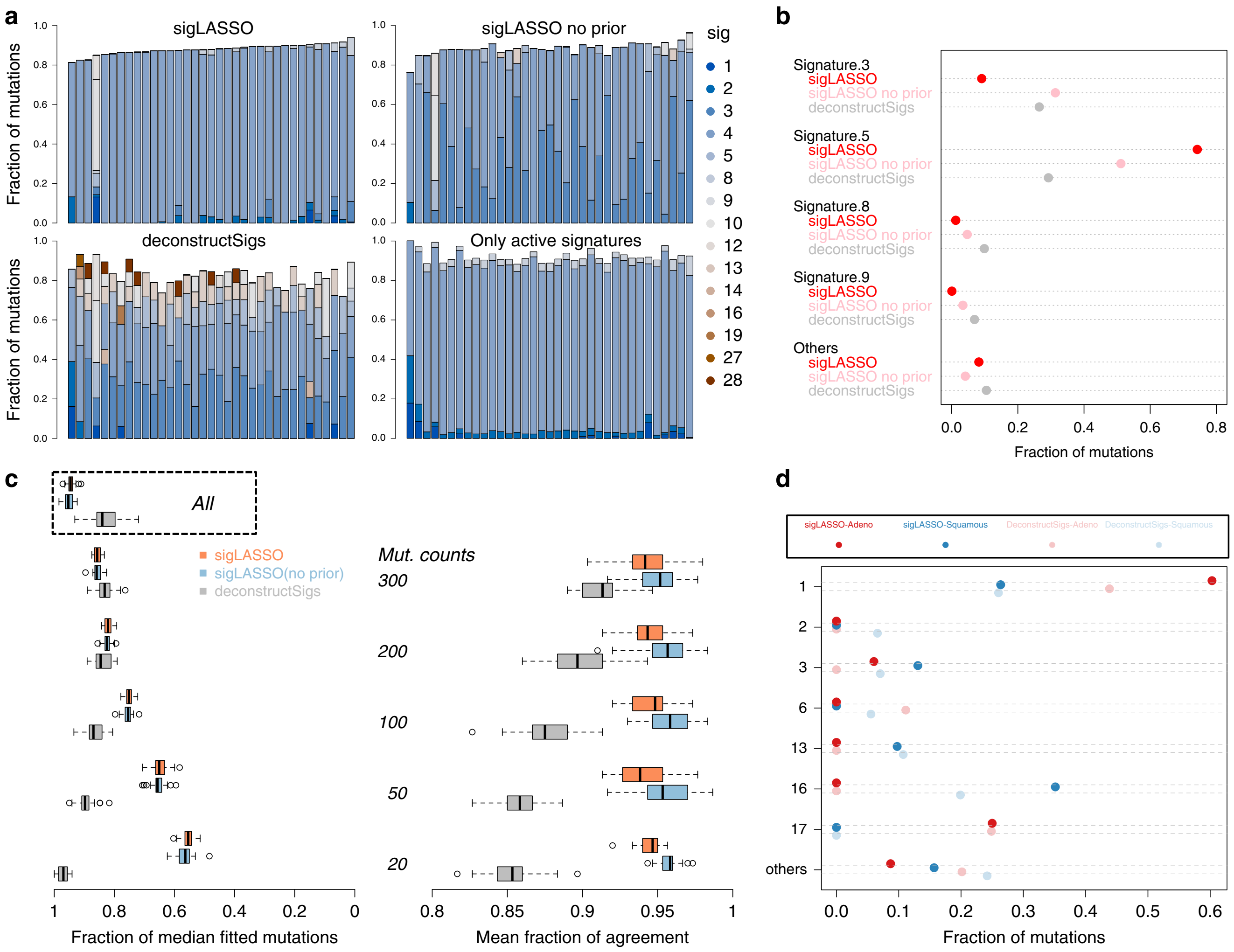

Fig. 3 Performance on pRCC and ESCA samples. a Signature assignment for 35 WGS pRCC samples. Bar plots show the fractions of mutation signature assignment for each sample using sigLASSO, sigLASSO without prior knowledge and deconstructSigs, and simple non-negative regression. **b** A dot chart showing the mean fraction of mutation signatures in each sample. Signatures that contributed <0.05 are grouped into others. **c** Subsampling (10 times each size, with no replacement) of 30 WGS pRCC samples that have more than 3000 mutations. The left panel shows the fraction of signature-fitted mutations. Results using all mutations are shown in box. The right panel shows the agreement (mean fraction of the consensus after binarizing whether a signature exists or not) among the methods. All $p \leq 2 \times 10^{-6}$ (paired two-sided Wilcox test between sigLASSO (and sigLASSO, no prior) and deconstructSigs). Box edges are the 25th and 75th percentiles and whiskers indicate the 1.5× IQR or max/min, which ones are smaller. **d** 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 <0.05 in all four cases are grouped into "others".

from zero, the performance immediately started improving and continued to do so. When more false signatures were mixed with true signatures given as prior knowledge, the performance slowly deteriorated. However, even with 1.5 times false signatures mixed in with true ones, the performance was slightly better (~2% in accuracy) than the baseline that used no priors. Stronger priors had larger boosting effects on the solution, as expected.

WGS scenario using renal cancer data sets. We next moved from synthetic data sets to real cancer mutational profiles, which are likely noisier than simulations and exhibit a highly non-random distribution of signatures.

We benchmarked the two methods using 35 WGS papillary kidney cancer samples[21]. The median mutation count was 4528 (range: 912–9257). We found that without prior knowledge, both sigLASSO and deconstructSigs showed high contributions from signatures 3 and 5 (Fig. 3a, b). deconstructSigs also assigned a high proportion to signatures 8(9.9%), 9(6.9%), and 16(4.7%). Signatures 3, 8, 9, and 16 were not found to be active in pRCC in previous studies and currently no biological support connects them to pRCC[2]. As expected, sigLASSO resulted in sparser solutions than deconstructSigs (mean signatures assigned: 3.40 and 4.43, respectively). Adding prior from COSMIC (kidney cancer combined) helps sigLASSO to put more weight on Signature 5, and increase the number of signatures assigned to 4.06.

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 ~10% mutations on average were assigned with other signatures. Moreover, the model assigned highly similar signature profiles to all samples. This finding suggests an overly simple, underfitted model.

To show that sigLASSO is sensitive to fitting uncertainty, able to abstain, and is robust, we performed subsampling on 30 pRCC samples (mutation counts ≥3000, Fig. 3c). As the sample size decreased, sigLASSO assigned fewer mutations with signatures (mean fraction dropped from 0.93 (all mutations) to 0.55 (20 mutations)), reflecting the greater uncertainty in fitting. This quality allows the user to be aware of the uncertainty in the solutions.

Moreover, the model complexity also declined accordingly (Supplementary Figure 3). The mean number of signatures assigned by sigLASSO decreased from 4.06 to 1.67, representing simpler models. Last, the performance of sigLASSO was robust and stable, as evidenced by stable outputs even in low sampling counts. Multiple subsampling runs showed high agreement in active and inactive signatures (~0.95 across all subsample sizes).

In contrast, deconstructSigs was unable to reflect model uncertainty. Surprisingly, as the mutation number decreased, the fraction of fitted mutations in deconstructSigs unexpectedly increased from 0.83 (all) to 0.97 (20 mutations). The model complexity, as reflected by the mean number of signatures assigned also increased from 4.43 to 4.70. The agreement between subsamples dropped significantly as the subsampling size decreased, indicating the solution is unstable and potentially overfitted.

WES scenario using esophageal carcinoma data sets. We next aimed to evaluate the two methods on 182 WES esophageal carcinoma (ESCA) samples with >20 mutations. The median mutation count was 175.5 (range: 28–2146), which is considerably lower than WGS but typical for WES.

In sigLASSO, the L1 penalty strength was tuned based on model performance. In a low mutation count setting, ESCA WES data set, the model variance was high, which pushed up the penalty. As expected by its ability to abstain, sigLASSO assigned a lower fraction of mutations with signatures than deconstructSigs (median: 0.68 and 0.90; interquartile range (IQR): 0.56–0.75 and 0.86–0.94, respectively, Supplementary Figure 4). deconstructSigs did not achieve 100% assignment because it performed a hard normalization of the coefficients after an unconstrained binary search and then discarded signatures that had ≤6% contribution. The leading signatures were 1, 16, 17, 3, and 13 in sigLASSO and 1, 17, 16, 6, 13, 3, and 2 in deconstructSigs.

deconstructSigs has been applied to distinguish between two different histological types of esophageal cancer[8]. We demonstrated that sigLASSO generates a sparser but comparable result with wider signature 1, and 16 gaps between the subtypes (Fig. 3d, Supplementary Figure 5). The adenocarcinoma subtypes had higher fractions of signature 1 and 17, and lower fractions of signature 3, 13, and 16.

Performance on 8893 TCGA samples. We ran sigLASSO and deconstructSigs with step-by-step set-ups on 8893 The Cancer Genome Atlas (TCGA) tumors (33 cancer types, Supplementary Table 11) with more than 20 mutations (Fig. 4, Supplementary Figure 6). The median mutation number is 100, IQR is 52–200.

Simple non-negative regression resulted in an overly dense matrix. Applying an L1 penalty made the solution remarkably sparse. Then, by incorporating the prior knowledge, the signature landscape further changed without significantly affecting the assignment sparsity. The change was inconsistent with the priors given. With low-count WES data (upper quartile: 200 and 96 mutational contexts), we expect the signature assignment to be very sparse as the uncertainty is high, the model should from making assignments. sigLASSO indeed assigned low number of signatures to the samples (median number of signatures assigned per sample: 1, IQR: 1–2). In comparison, the solutions of deconstructSigs were less sparse (median number of signatures assigned per sample: 4, IQR: 3–5). In all, 18.5% samples were fitted by deconstuctSigs with six signatures or more, compared with 0.4% by sigLASSO. We provide a kidney cancer example and detailed paired analysis in the Supplement (Supplementary Figure 7, 8).

When the mutation number cutoff increases to 50, the above results of the two methods still stand (Supplementary Figure 9). We also released all of the signature results of 8893 TCGA samples on the sigLASSO GitHub site.

Using the large-scale tumor signature profiles, we further explored the correlation of smoking signature and smoking status using annotations from a previous study[15]. In lung adenocarcinoma (LUAD) samples, smoking samples carry significantly higher signature 4 (“smoking signatures”) fractions than nonsmoking ones (median: 0 and 0.68, respectively, $p \le 1 \times 10^{-15}$, Wilcoxon rank test, Supplementary Figure 10). Similarly, in lung squamous cell carcinoma (LUSC) samples, we observed high fractions of signature 4 in smokers, but because only 3.5% of the LUSC cohort are nonsmokers (6/171), we were underpowered to draw a statistical significance. In non-lung cancer samples ($N = 1500$), we also found a weaker but statistically significant trend of higher signature 4 in smokers (mean: 0.008 and 0.038, respectively, $p = 3.3 \times 10^{-8}$, Wilcoxon rank test, Supplementary Figure 10). This result is in agreement with previous studies on smoking signatures[15].

sigLASSO is computationally efficient. sigLASSO iteratively solves two convex problems. The $\vec{w}-$step can be solved using a very efficient coordinate descent algorithm (glmnet)[14]. The $\vec{p}-$step is solved by a set of quadric equations. We observed empirically that the solution quickly converges in a few iterations even with low mutation numbers. Meanwhile, deconstructSigs uses binary search instead of regression to try every coefficient 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 remarkably affected the running time of sigLASSO. With a high mutation number, sigLASSO was 3–4 times faster than deconstructSigs; with a low mutation number (50 mutations), these two tools showed a comparable computation time. Noticeably, despite using less time sigLASSO employs empirical parameterization and alternative optimization, which regresses the signature-fitting problems hundreds of thousands of times with different parameters in a typical run. Therefore, by carefully designing an effective algorithm, the core fitting step of sigLASSO is orders of magnitudes faster than deconstructSigs. This enables sigLASSO to probe the data complexity and accordingly tune the model complexity.

Discussion

Studies decomposing cancer mutations into a linear combination of signatures have provided invaluable insights into cancer biology[4–6,22]. Indeed, researchers have gained a better understanding of one of the fundamental driving forces of cancer initiation and development, mutagenesis, by inferring mutational signatures and latent mutational processes.

A practical problem for many researchers is how to leverage results from large-scale signature studies and apply them to a small set of incoming samples. Although this might seem to be a simple linear system problem, core challenges include (1) preventing over- and underfitting on only one sample, often with very few mutations (especially in WES), (2) achieving proper variable selection from a large amount of recognized signatures, and (3) promoting interpretability.

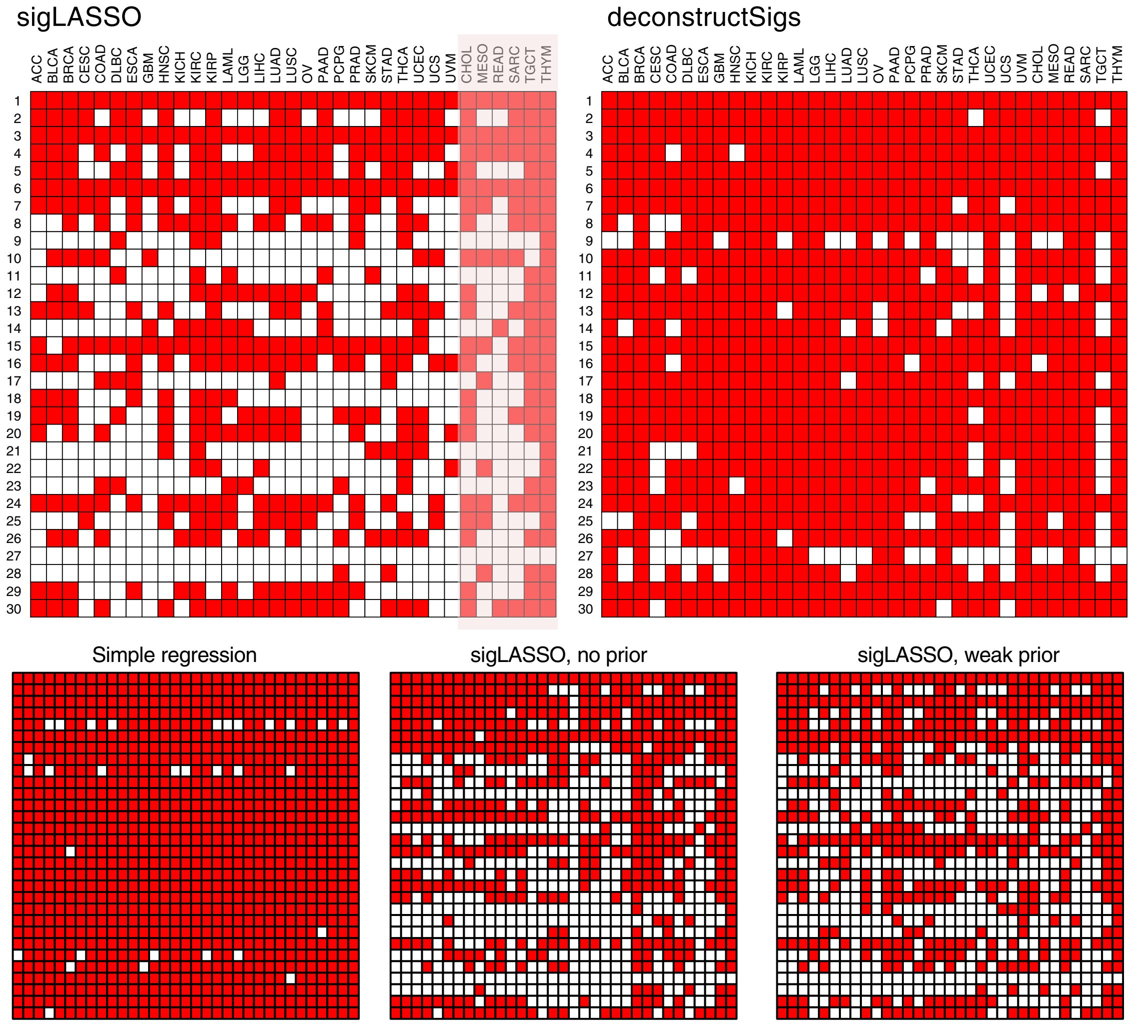

Fig. 4 Performance on 33 TCGA cancers. Active signatures (total contribution >0.1%) in 33 cancer types using different methods. Only 27 cancer types have previously known signature distributions (others are shaded). Penalty coefficients of priors used for sigLASSO: 0.01. Weak priors: 0.5. Priors were taken from COSMIC, and provided with our "sigLASSO" R package.

First, with less than a few hundred mutations, sampling variance becomes a significant factor in reliably identifying signatures. 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 large number of available signatures necessitates a proper variable selection, especially with a limited amount of observed data. Signature studies on large-scale cancer data sets 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 2–13 known signatures are observed in a given cancer type (based on 30 COSMIC signatures), which might include hundreds and even thousands of samples. Therefore, sparse solutions are in line with previous de novo studies, and are biologically sound and interpretable. In addition, sparse solutions can give better predictions owing to lower estimator variance. Third, a desirable method should be parameterized according to the data complexity to achieve optimum fitting. Finally, mutational signatures are not orthogonal owing to their biological nature. Collinearity of the signatures will lead to unstable fittings that change erratically with even a slight perturbation of the observation.

deconstructSigs was the first tool that could identify signatures even in a single tumor. Instead of regression, this tool uses binary search to iteratively tune coefficients. To achieve sparsity, deconstructSigs performs post hoc pruning with a preset 6% cutoff value. The mutation spectrum is normalized before fitting, thus making mutation counts invariant to the model. Moreover, the binary search operates on unbounded coefficients and uses a hardcoded upper and lower bounds, which provides no guarantee of finding the optimal solution. Finally, the greedy nature of stepwise coefficient tuning is prone to eliminating valuable predictors in later steps that are correlated with previously selected ones[23].

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 achieving effective variable selection and promoting sparsity. 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

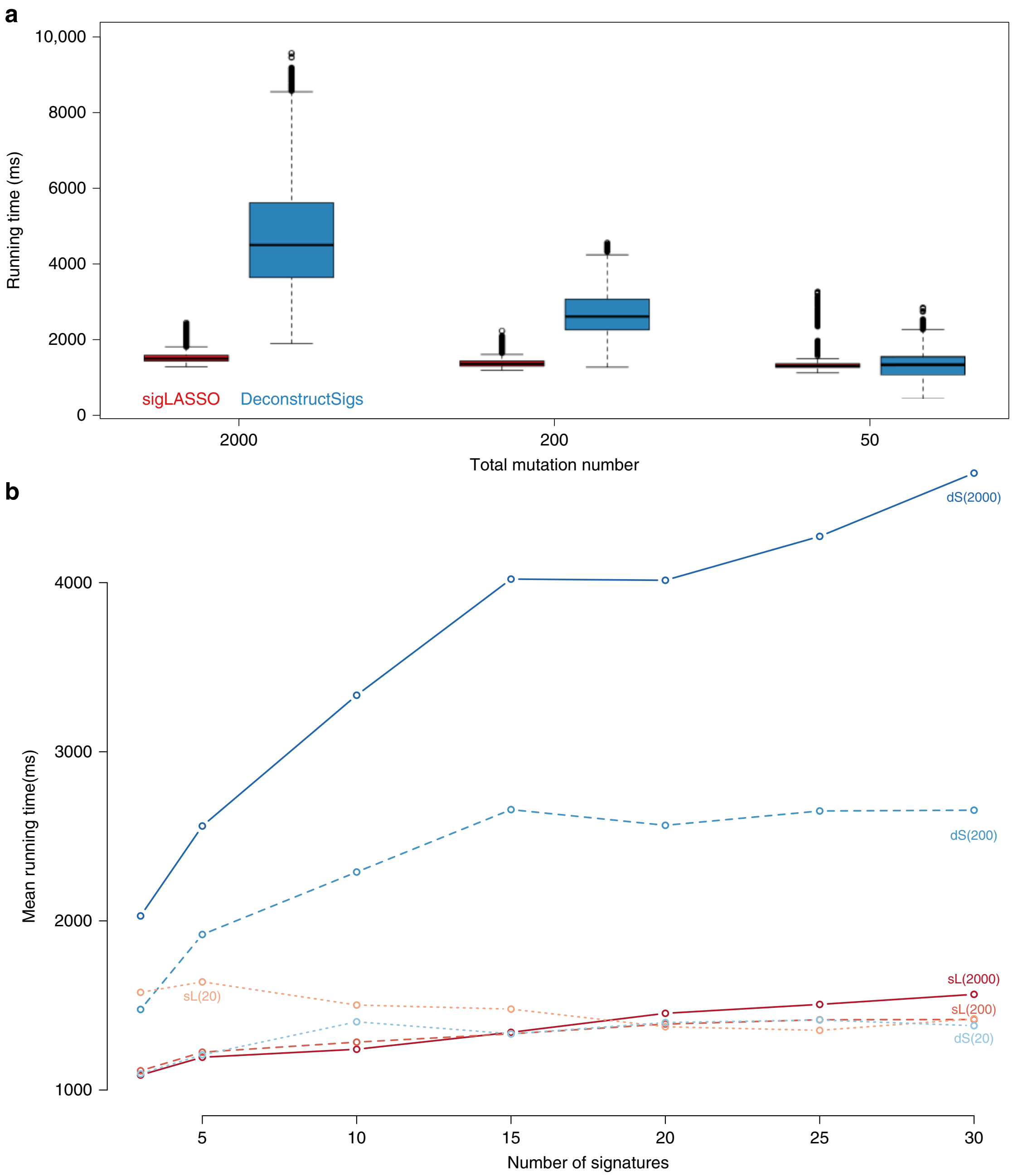

Fig. 5 sigLASSO is computationally efficient. a Running time of sigLASSO and deconstructSigs at different total mutations numbers. Box edges are the 25th and 75th percentiles and whiskers indicate the 1.5× IQR or max/min, which ones are smaller. **b** Running time of sigLASSO at different numbers of signatures (downsampled from 30 COSMIC signatures) simulated with three different mutations numbers (in parenthesis). Noise level: 0.1 sL: sigLASSO, dS: deconstructSigs.

corresponding distribution, we demonstrated that sigLASSO achieves better and more stable signature attribution, especially in cases with low mutation counts. In cancer research, WES data are abundant, but it also suffers from undersampling in signature attribution. In these cases, sigLASSO generates more reliable and robust solutions. We showcase the successful application of sigLASSO on 8893 TCGA WES samples. Overall, sigLASSO achieves better sparsity than deconstructSigs and promote structures in the previous signature studies by injecting priors into the optimization process. The final signature distribution of neither sigLASSO nor deconstructSigs is identical to the original signature studies, which is probably owing to biases in the original signature discovery, sequencing/variants calling, and unknown hypermutational processes. We noticed deconstructSigs assigned signatures to 100% of the mutations in most samples, while in sigLASSO assignment fraction correlated with cancer types. This abstain pattern indicates the presence of potentially unknown signatures and cancer subtypes.

As the cost of WGS drops rapidly, we expect an even greater number of cancer samples to be sequenced[24]. 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 regularization to achieve a simpler, sparser, and more stable solution. Furthermore, sigLASSO is very efficient and thus able to handle a larger number of samples and signatures, as well as hyperparameter optimization.

Moreover, sigLASSO does not specify a noise level explicitly beforehand, but instead empirically tunes parameters based on model performance. By contrast, deconstructSigs specifies a noise level of 0.05 to derive a cutoff 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. Abstention is a desired trait in machine learning models. In addition to the objectives, the model learns what it knows and what it does not (KWIK: Knows What It Knows),[25] In both subsampling and simulation experiments, sigLASSO is able to abstain under high uncertainty, providing solutions with consistently high precision and robustness under various scenarios.

Next, owing to the collinearity nature of signatures, pure mathematical optimization might lead algorithms 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 input could be cancer type-specific signatures or patient clinical information (e.g., smoking history or chemotherapy). Furthermore, we transformed the binary classification (active or inactive) to a continuous space, with weights indicating the strength of the prior. These weights could be effectively learned using patient information or other assays (e.g., RNA sequencing or methylation arrays). Moreover, sigLASSO can adapt to sparsity promoting schemes in de novo discovery tools through priors. For example, SparseSignatures[12] also uses L1 regularization but does not penalize on background signatures. sigLASSO could use this piece of information in its priors to better assign signatures when working together with SparseSignatures. We showcased the performance of sigLASSO on real cancer data sets. 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.

We also implemented elastic net with hyperparameter optimization in sigLASSO. Elastic net blends L2 with L1 regularization and in principle demonstrates better performance and stability than LASSO on strongly correlated features[26]. We found that elastic net did not improve the performance (measured as meaningful reduction in MSE in cross-validation) in our simulations using the current 30 COSMIC signatures, likely because the correlations between them are not too high. However, with more cancer samples sequenced every day, researchers will gain power to discern highly correlated mutational processes and grow the size of the signature set significantly. Therefore, elastic net might be beneficial in the near future.

Finally, sigLASSO uses quadratic loss, which follows the previous de novo studies[2,3,5]. By doing so, sigLASSO is unbiased in detecting signatures in cancer samples. Nonetheless, the initial discovery suffers from several limitations. First, the actual number of mutational processes is likely higher than 30, the current number of COSMIC signatures. This is because the power of the de novo discovery is bounded by the amount of available data. Second, the nature of mutagenesis leads to some mutational processes being more prevalent than others. Some of the signatures are cancer type specific. The original signature discovery did not use a balanced data set with equal numbers of cancer types. Third, the original NMF objective function employs quadratic loss, which might lead to bias towards flat signatures over spiky ones. In addition to the original NMF approach, researchers have proposed other decomposition methods for signature discovery. For example, SomaticSignatures[27] uses PCA. Because the loss function is also quadratic, we expect sigLASSO to work seamlessly with SomaticSignatures.

Our simulation reveals that sigLASSO does not have a discernable preference to assign to either spiky or smooth signatures (Supplementary Figure 12). However, we found that the signature distribution from sigLASSO and deconstructSigs cannot be fully explained by the current signature knowledge. Further research probing the biological foundation of the signatures and quantifying biological priors will help advance our understanding and improve the interpretability of signature assignments. Meanwhile, as there are still many unsolved issues in signature discovery, we advocate that sigLASSO should also be used with discretion and its results interpreted cautiously.

sigLASSO exploits the previously overlooked mutation sampling uncertainty and formulates a framework that jointly optimizes the objective (i.e., signature fitting with L1 regularization) and the sampling likelihood. In biological experiments, low-count observations on discrete variables are common. For example, in single cell RNA sequencing (scRNAseq), the measured discrete mRNA counts are often very low or even zero (owing to undersampling). Our joint optimization approach could have further implications in these scenarios. We, indeed, find some similar work in scRNAseq[28,29]. Another method for de novo signature discover, EMu, formulates the discreet mutational process as a Poisson generative model[30]. When the total mutational count is fixed, such a Poisson generative model is equivalent to our multinomial sampling process. Our work also could be extended to mutagenesis modeling and parameter estimation—for example, in estimating the nucleotide-specific background mutation rate in cancer.

Methods

Optimizing sigLASSO. The negative log-likelihood is convex in respect to both $\vec{p}$ and $\vec{w}$ 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 using alternative convex search, which iteratively updates these two variables[31].

Algorithm 1.

sigLASSO algorithm

```
initialization: $p_i^0 \leftarrow p_i^{mle} = \frac{m_i}{\sum_{i=1}^{n} m_i}$;   $t \leftarrow 0$
while $t < t_{max}$ do
    $\vec{w}^{t+1} \leftarrow \mathrm{argmax}\, \frac{\alpha}{2} \sum_{i=1}^{n} \left( p_i^t - \sum_{k=1}^{K} s_{ik} w_k \right)^2 - \lambda \sum_{k=1}^{K} c_k w_k$   ($\vec{w}$-step)
    $\vec{p}^{t+1} \leftarrow \mathrm{argmax} \sum_{i=1}^{n} \left\{ m_i \log p_i - \frac{\alpha}{2} \left( p_i - \sum_{k=1}^{K} s_{ik} w_k^{t+1} \right)^2 \right\}$   ($\vec{p}$-step)
    if$\| \vec{p}^{t+1} - \vec{p}^t \|_2^2 < \epsilon$ then
        break
    $t \leftarrow t+1$
return $\vec{w}^{t+1}$
```

Specifically, to begin the iteration, we initialized $\vec{p}$ using MLE. We started with the $\vec{w}-$step, which is a non-negative linear LASSO regression that can be efficiently solved by glmnet.[14] λ is parameterized empirically.

Next, we solved the $\vec{p}$ with a Lagrange multiplier to maintain the linear summation constraint $\sum_{i=1}^{n} p_i = 1$. The non-negative constraint of $_{pi}$ is satisfied by only retaining the non-negative root of the solution.

Intuitively, in the $\vec{p}-$step, we tried to estimate $\vec{p}$ by optimizing the multinomial likelihood while constraining it to be not too far away from the fitted $\vec{p}$. If we only used the point MLE of $\vec{p}$ based on sampling and did not perform the $\vec{p}-$step, the model would assume the sampling is perfect and become insensitive to the total mutation counts. The trade-off in the $\vec{p}-$step between the multinomial likelihood and the square 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 $\vec{p}$ all affect the tension between sampling and linear fitting.

Optimizing the $\overrightarrow{p}$-step. In the $\vec{p}$-step, we tried to solve the following problem with $\tilde{p}_i$ from the $\vec{w}-$step.

$$\begin{aligned} \vec{p} &= \operatorname{argmax} \sum_{i=1}^{n} \{m_i \log \mathrm{p_i} - \tfrac{\alpha}{2}(\mathrm{p_i} - \tilde{\mathrm{p}}_\mathrm{i})^2\} \\ \text{s.t.} \forall p_i &\geq 0, \sum_{i=1}^{n} p_i = 1, \tilde{p}_i = \sum_{k=1}^{K} s_{ik} w_k \end{aligned} \tag{4}$$

We added the Lagrangian multiplier Λ to satisfy the linear constraint of $\sum_{i=1}^{n} p_i = 1$ and took the derivatives with respect to p_i ($i = 1, 2. . . n$) and Λ. This resulted in $n + 1$ equations.

$$\begin{gathered} p_1^2 - \left(\tfrac{\Lambda}{\alpha} + \tilde{p_1}\right)p_1 - \tfrac{m_1}{\alpha} = 0 \\ \ldots \\ p_i^2 - \left(\tfrac{\Lambda}{\alpha} + \tilde{p_i}\right)p_i - \tfrac{m_i}{\alpha} = 0 \\ \ldots \\ p_n^2 - \left(\tfrac{\Lambda}{\alpha} + \tilde{p_n}\right)p_n - \tfrac{m_n}{\alpha} = 0 \\ \sum_{i=1}^{n} p_i = 1 \end{gathered} \tag{5}$$

The roots of the first n quadratic equations are given by

$$p_i = \frac{\left(\tilde{p_i} + \frac{\Lambda}{\alpha}\right) \pm \sqrt{\left(\tilde{p_i} + \frac{\Lambda}{\alpha}\right)^2 + 4\frac{m_i}{\alpha}}}{2} \tag{6}$$

$\alpha = 1/\sigma^2$ is strictly positive and m_i is non-negative. Therefore, if $m_i = 0$, there exists only one zero root and $p_i = 0$ iff. $m_i = 0$. If $m_i > 0$, there is exactly one negative and one positive root. Because we required $\forall p_i \geq 0$, we only kept 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 non-negative maximum.

We plugged all the roots into the last equation (i.e., the linear constraint) and used the R function uniroot() to solve Λ.

Parameter tuning. We tuned λ 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 often correlated, we split the data set into eight subsets. Each subset contained two of each single-nucleotide substitution. We then held one subset as the testing data set 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 an MSE 0.5 or 1 SD from the minimum MSE. λ was tuned whenever $\vec{p}$ deviated far from the estimation from the previous round. By adaptively learning $\vec{p}$, 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 affecting convergence.

$\alpha = 1/\sigma^2$, σ^2 is estimated using $\sigma^2 = \frac{SSE}{(n-k)}$, where k is the number of non-zero coefficients in the LASSO estimator and SSE is the sum of squared errors[32]. sigLASSO updates α after every LASSO linear fitting step. To avoid grossly overestimating σ^2 (thus underestimating α) in the initial steps when $\vec{p}$ is far from the optimum, we set a minimum α value. In addition, because in practice factors such as unknown signatures violate the assumptions of linear signature regression, α tends to get overestimated. So we multiplied it by a confidence factor that represents how confident we are about the goodness of the signature fit. By default, we set min $\alpha =$ 400 and the confidence factor to 0.1. Users can further tune these values based on the strength of prior belief of noise level and confidence level of the signature model.

Option for elastic net. Adding an L2 regularizer (i.e., elastic net) might improve and stabilize the performance when variables are highly correlated[26]. Therefore, we also implemented elastic net in sigLASSO. The objective function then became:

$$\ell = \sum_{i=1}^{n} \left\{ m_i \log \mathrm{p_i} - \frac{\alpha}{2} \left(\mathrm{p_i} - \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{s_{ik} w_k} \right)^2 \right\} - \lambda \sum_{k=1}^{K} c_k (\gamma w_k + (1 - \gamma) w_k^2) \tag{7}$$

We tuned the hyperparameter γ by grid search together with λ. We always picked the largest γ and the largest λ (for better sparsity) that gave an MSE that was 0.5 SD from the minimal MSE. In simulations, we did not find that introducing L2 regularization added additional benefits. The model gave almost identical solutions to using L1 only. Nonetheless, we kept elastic net as an option in highly correlated signature scenarios to the user in our implementation.

Data simulation and model evaluation. We downloaded 30 previously identified COSMIC signatures v2 (http://cancer.sanger.ac.uk/cosmic/signatures). We created a simulated data set by randomly and uniformly drawing two to eight signatures and corresponding weights (minimum: 0.02). The reason for picking up to eight signatures is because (1) empirically, in cancer signature studies, most samples have only a few signatures. (2) We further confirmed the sparsity of signatures by running deconstructSigs on large-scale TCGA data and found 99.96% samples got assigned with eight or less signatures. (3) Moreover, biologically, we believe that it is unreasonable for more than eight of the processes to act simultaneously in one sample. Many of these processes describe biological disjoint conditions, e.g., smoke, UV radiation and tissue-specific cellular processes. We simulated additive Gaussian noise at various levels with a positive normal distribution of up to 25 (1–25, uniformly drawn) randomly selected 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, both with and without prior knowledge of the underlying signature. To evaluate their performance, 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 sigLASSO on 8893 TCGA samples. To assess the performance of our method on real-world cancer data sets, we used somatic mutations from 33 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 Supplementary Table 1.

We extracted prior knowledge on active signatures in various cancer types from a previous pan-cancer signature analysis (http://cancer.sanger.ac.uk/cosmic/signatures).

sigLASSO software suite. sigLASSO is implemented as an R package “siglasso”. It accepts processed mutational spectrums and VCF files. We provided functions to parse mutational spectrums from VCF files as well as some visualization methods in the package. “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. It is computationally efficient; using default settings, the program can successfully decompose a whole genome sequencing (WGS) cancer sample in less than a few seconds on a regular laptop. For time profiling purposes, 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 100 noiseless simulated data sets and repeated the process 100 times for each evaluation.

We made siglasso source code available at http://github.com/gersteinlab/siglasso.

Evaluation criteria for signature assignment. 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 have relied largely on mutagen exposure association from patient records and biochemical knowledge on mutagenesis. Here, besides using simulation, we illustrated the outputs of different models and compared the results on real data set with existing signature knowledge and distributions. 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. A sparse distributed signature set among cancer samples and types is defined in previous de novo discovery studies. Any signature-fitting tool should follow and produce a similar signature distribution. Moreover, 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. Existing signature-identifying methods aim to implicitly achieve sparse solutions by dropping signatures with small coefficients or pre-selecting a small signature subset for fitting.

The ability to abstain (decline to assign signatures to all mutations): signature assignment issues are often undefined owing to collinearity of the signatures and a larger number of possible signatures. Overfitting is a serious concern, especially when the mutation counts are low (e.g., WXS or cancer genomes with fewer mutations) or the fitting is poor (e.g., unknown mutational process or sample contamination causing high noise). A good solution should refuse to fully assign signatures to every mutation when it does not have enough confidence to do so and, instead, defer to the human researchers to decide.

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 owing to undersampling. In particular, under low mutation count, not all of the operative signatures would be reliably discovered. It is better to abstain under high uncertainty.

Biological interpretability: the solution should be biological interpretable. Because of the biological nature of collinearity in the signatures, simple mathematical optimization might pick the wrong signature. Researchers now tackle 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, leaving space for noise and rare or unknown signatures. For instance, we expected to find divergent signature distributions in different cancer types, which is reported in previous de novo discovery studies. Various tissues have divergent endogenous biological features, are exposed to diverse mutagens, and undergo mutagenesis in dissimilar fashions. Signature patterns should be able to distinguish between cancer types. This can be achieved by using different cancer type-specific priors.

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

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are avaiable from Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). A detailed list of files used in this study can be found in Supplementary Table 1.

Code availability

We make the code publicly available at github.com/gersteinlab/siglasso.

Received: 7 January 2020; Accepted: 22 June 2020;
Published online: 17 July 2020

References

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 Rep.* **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. *Nat. Rev. Genet.* **15**, 585 (2014).
5. Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* **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. *bioRxiv036541* (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 Biol.* **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. *Nat. Commun.* **6**, 8683 (2015).
10. Rahbari, R. et al. Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126 (2016).
11. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
12. Ramazzotti, D., Lal, A., Liu, K., Tibshirani, R. & Sidow, A. De novo mutational signature discovery in tumor genomes using sparsesignatures. *bioRxiv384834* (2018).
13. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **58**, 267–288 (1996).
14. Friedman, J., Hastie, T. & Tibshirani, R. *glmnet: Lasso and elastic-net regularized generalized linear models.* R package version1 (2009).
15. Alexandrov, L. B. et al. Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618–622 (2016).
16. Viel, A. et al. A specific mutational signature associated with dna 8-oxoguanine persistence in mutyh-defective colorectal cancer. *EBioMedicine* **20**, 39–49 (2017).
17. Schulze, K. et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat. Genet.* **47**, 505 (2015).
18. Davies, H. et al. Hrdetect is a predictor of brca1 and brca2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517 (2017).
19. Park, T. & Casella, G. The bayesian lasso. *J. Am. Stat. Assoc.* **103**, 681–686 (2008).
20. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **101**, 1418–1429 (2006).
21. Li, S., Shuch, B. M. & Gerstein, M. B. Whole-genome analysis of papillary kidney cancer finds significant noncoding alterations. *PLoS Genet.* **13**, e1006685 (2017).
22. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385 (2018).
23. Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. et al. Least angle regression. *Ann. Stat.* **32**, 407–499 (2004).
24. Muir, P. et al. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.* **17**, 53 (2016).
25. Li, L., Littman, M. L., Walsh, T. J. & Strehl, A. L. Knows what it knows: a framework for self-aware learning. *Mach. Learn.* **82**, 399–443 (2011).
26. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **67**, 301–320 (2005).
27. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. Somaticsignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**, 3673–3675 (2015).
28. Qiu, X. et al. Single-cell mrna quantification and differential analysis with census. *Nat. Methods* **14**, 309 (2017).
29. Zhu, L., Lei, J., Devlin, B. & Roeder, K. A unified statistical framework for single cell and bulk rna sequencing data. *Ann. Appl. Stat.* **12**, 609 (2018).
30. Fischer, A., Illingworth, C. J., Campbell, P. J. & Mustonen, V. Emu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol.* **14**, R39 (2013).
31. Gorski, J., Pfeuffer, F. & Klamroth, K. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Math. Methods Oper. Res.* **66**, 373–407 (2007).
32. Reid, S., Tibshirani, R. & Friedman, J. A study of error variance estimation in lasso regression. *Stat. Sin.* **26**, 35–67 (2016).

Acknowledgements

This work was supported in part by NIH/NHGRI (1R01HG008126), NIH/NICHD (1DP2HD091799-01), A.L. Williams Professorship and the facilities and staffs of the Yale University Faculty of Arts and Sciences High Performance Computing Center. We thank Harrison Zhou, Cong Liang, Jinrui Xu, Diego Galeano, and Mengting Gu for helpful discussion.

Author contributions

S.L. and M.G. conceived and designed research. S.L. designed the algorithm, developed the software, and carried out experiments. S.L. and F.W.C. performed the theoretical analysis of the algorithm. S.L., F.W.C., and M.G. analyzed the results. S.L. and M.G wrote the manuscript. M.G. supervised the study.

Competing Interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-17388-x.

Correspondence and requests for materials should be addressed to M.B.G.

Peer review information *Nature Communications* thanks Javier Herrero and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.