SigLASSO: a LASSO regression approach for identifying mutational signatures in cancer genomics

Shantao 2/19/2018 7:10 PM
Deleted: identification

Abstract: Multiple mutational processes fuel carcinogenesis and leave characteristic signatures in cancer genomes. Identifying operative mutational processes by signatures helps understand cancer initiation and development. The task is to delineating cancer mutations by nucleotide context into a linear combination of mutational signatures. Past mutational signature studies suggest the solution should be sparse to be biologically interpretable. Previously published methods use empirical forward selection or iterate signature combinations by brutal force. Here, we *alternatively* formulate the problem as a LASSO linear regression and accordingly developed a software tool, SigLASSO. By parsimoniously assigning signatures to cancer genome mutation profiles, the solution becomes sparse and more biologically interpretable. Additionally, SigLASSO integrates biological prior knowledge into the solution by fine-tuning penalties on coefficients. Compared with subseting signatures before fitting, our method leaves leeway for noises and unknown signatures. Last, the model complexity is informed by the size and complexity of the data through parameterizing using cross-validation and subsampling.

Shantao 2/19/2018 6:37 PM
Deleted: using

Shantao 2/2/2018 8:51 PM
Deleted: organically

Shantao 2/19/2018 6:38 PM
Deleted: by

Introduction

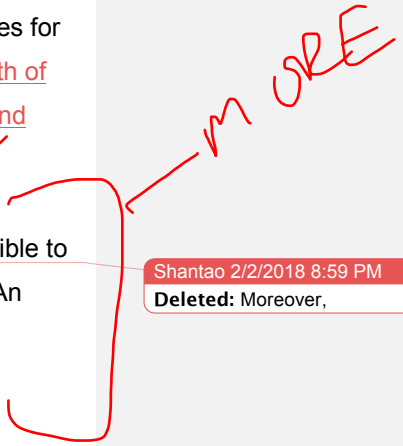
Mutagenesis is the fundamental process for cancer development. Examples include spontaneous deamination of cytosine, ultraviolet light inducing pyrimidine dimer and alkylating agents crosslinking guanines. Multiple endogenous and exogenous mutational processes drive cancer mutagenesis and leave distinct fingerprints. Noticeably, these processes have characteristic mutational nucleotide context biases. Mutation profiling of cancer sample at manifestation finds all mutations accumulate over lifetime, including somatic alterations both

before cancer initiation and during cancer development. In a generative model, over time multiple latent processes generate mutations drawing from their corresponding nucleotide context distributions ("mutation signature"). In cancer samples, mutations from various mutation processes are mixed and observable by sequencing.

Applying unsupervised methods such as non-negative matrix factorization (NMF) and clustering to large-scale cancer studies, researchers have identified at least 30 mutational processes [REF]. Many processes are recognized and linked with known etiologies, for example aging, smoking or ApoBEC activity. Investigating the fundamental underlying processes helps understand cancer initiation and development.

Shantao 2/19/2018 8:46 PM
Deleted: thirty

One prominent task in nowadays cancer research is to leverage on signatures studies on large-scale cancer cohorts and efficiently assign active signatures for new cancer samples [REF]. Although scientists do not have the ground truth of mutational signatures in cancer samples, they do have some reasonable and logical expectations about the solution. In this work, we aim to design a computational framework to achieve these expectations. For example, we believe the solution should be sparse as past studies indicate it is not possible to have all signatures active in a single sample or even a given cancer type. An apparent example is, UV-associated signatures should not be observed in tissues that are not exposed.



Shantao 2/2/2018 8:59 PM
Deleted: Moreover,

Previously published methods use forward selection with an empirical stopping criterion or iterate all combinations (brutal force). Here, we alternatively formulate it as a more mathematically rigorous LASSO linear regression problem. Out approach is the first one that explicitly penalizes the model complexity in optimization. We use L1 norm as the regularizer as L0 norm (cardinality of active signatures) is designed but cannot be effectively optimized. L2 norm, on the

other hand, leads to many small, non-zero coefficients. By penalizing the L1 norm of coefficients, the algorithm is efficient and produces sparse and biologically interpretable solutions. Additionally, this approach is able to organically integrate biological prior knowledge into the solution by fine-tuning penalties on the coefficients. Compared with current approach of subseting signatures before fitting, our soft prior method leaves leeway for noises and unidentified signatures. Last, unlike previous methods, SigLASSO is aware of data complexity such as mutational number and patterns. Our method is automatically parameterized based on cross-validation and subsampling, allowing data complexity to inform model complexity. This approach promotes results replicability and fair comparison across datasets.

Material and methods

Signature identification problem

Different mutational processes leave mutations in the genome with distinct nucleotide contexts. In particular, we consider the mutant nucleotide context and look one nucleotide ahead and behind. This divides mutations into 96 trinucleotide contexts. Each mutational process carries its unique signature, which is represented by a mutational trinucleotide context distribution (Fig 1A). 30 signatures are identified by nonnegative matrix factorization (NMF) and clustering from large-scale pan cancer analysis (REF). Here our objective is to leverage on the pan cancer analysis and decompose mutations observed in new samples into a linear combination of signatures. Mathematically, the problem is formulated as the following nonnegative regression problem:

$$\min_{W \in R^+} \|SW - M\|_2$$

The mutation matrix, *M*, contains mutations of each sample broken down into 96 nucleotide contexts. *S* is a 96×30 signature matrix, containing the mutation probability in 96 trinucleotide contexts of the 30 signatures. *W* is the weights matrix, representing the contributions of 30 signatures in each sample.

SigLASSO workflow

To promote sparsity and interpretability of the solution, SigLASSO uses LASSO regression, adding an L1 norm regularizer on the weights (i.e. coefficients) of the signatures. LASSO is mathematically justified and can be computationally efficiently solved by using least-angle regression (REF). Mathematically, LASSO is equivalent to a Bayesian linear regression framework with Laplace prior.

$$\min_{W \in R^{+}} (\|SW - M\|_2 + \sum \lambda I \|W\|)$$

λ is parameterized by 10-fold cross validation. We use the smallest λ that gives mean square error (MSE) within 3 standard deviances (SD) of the minimum.

Mutation count is an important factor affecting signature identification. To assess the solution stability and adjust for lower signature ascertainment when fewer mutations are observed, SigLASSO performs subsampling. At each subsampling step, it samples 50% mutations, solves the regression problem and finds active (i.e. with nonnegative coefficients) signatures. In the end, we only retain signatures that are active in more than τ fraction of all subsampling trials. τ can be set empirically between 0.6 to 0.9 (REF). In our study, we use 0.6 and set subsampling to 100 times unless otherwise specified.
A schematic illustration of the SigLASSO workflow is shown here (Fig 1B).

Fig1: A: Mutational processes have different mutational contextual spectrums (mutational signature) and contribute with different weights (loadings) to the final observable mutation spectrum in cancer. **B:** A schematic illustration of SigLASSO workflow.

Data simulation and model evaluation

First we downloaded 30 previously identified signatures (http://cancer.sanger.ac.uk/cosmic/signatures, REF). We created simulated dataset by randomly and uniformly drawing signatures (2 to 8 signatures) and

corresponding weights (minimum: 0.02). Noise was simulated at various levels with a uniform distribution on 96 trinucleotide contexts. Then we summed up all the signatures and noise to form a mutation distribution. We randomly drew mutations from this distribution with different mutation counts.
We ran deconstructSigs according to the original publication (REF). To evaluate the performances, we compared the inferred signature distribution with the simulated distribution and calculated mean square error (MSE). We also measured the number of false positive signatures in the solution as well as the false negative ones.

Illustrating on real dataset

To assess the performance of our method on real world cancer dataset, we use TCGA somatic mutations from various cancer types. VCF files are downloaded from Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). A detailed list of files used in this study can be found in Appendix X.
The signature composition results were compared with previous pan cancer signature analysis (http://cancer.sanger.ac.uk/cosmic/signatures, REF). Priors used in SigLASSO were also extracted from this source.

SigLASSO software suite

SigLASSO accepts (vcf files or) processed mutational spectrums. It allows the users to specify biological priors, subsampling steps and subsampling cutoff. SigLASSO uses the 30 COSMIC signatures by default. Users are given the option to also supply customized signature files. LASSO is computationally efficient. Using default settings, the program could successfully decompose a cancer sample data in a few seconds on a regular laptop (3 GHz i7 CPU, 16 GB DDR3 memory).
SigLASSO is released as an R package (SigLASSO). Updated code is also distributed on GitHub (https://github.com/ShantaoL/SigLASSO).

Results

1. Performance on simulated dataset

Both SigLASSO and deconstructSigs perform better with higher mutation number and lower noise (Fig 2). In general, the MSE is below 0.02 with high mutations and low noise (0.1). This performance is remarkably good for both programs. Even a program that recovers all signatures perfectly but also oblivious about the noise, MSE will be the square of noise level, which is 0.01 in this case. Likewise, MSE should be 0.04 when noise level rises to 0.2. And this is what we observe generally in both programs.

Fig2: Performance of sigLASSO and deconstructSigs in four different scenarios, with high/low noise and high/low mutation counts. Error bars indicate one standard deviation (SD) of ten repeats.

When mutation number decreases, we introduce uncertainty in sampling, which is negligible in high mutation number cases. As expected, the MSE jumped into the 0.1 to 0.3 range for both low and high noise. Clearly, the error here is dominated by undersampling, not the noise we embedded.

[[Also want to do a simulation to show benchmark on individual signatures, and how prior helps to improve performance]]

2. Performance on real dataset

Then we moved from synthetic datasets to real cancer mutational profiles. One of the problem in cancer signature research is the ground truth of real samples cannot be obtained. Previous large-scale signature studies largely replies on mutagen exposure association from patient records and biochemistry expertise about mutagenesis. Here, we illustrated the outputs of different models and simply compared the results with existing signature knowledge.

Although there is no golden standard to evaluate the performance, we do have a few reasonable expectations about the solution.

1) One or more signatures should be active in a given cancer sample and type. However, not all signatures are active in a given cancer sample or type. An obvious example is the UV signature should not be observed in tissues unexposed.

2) We expected to find divergent signature distributions in different cancer types. Various tissues are exposed to diverse mutagens and undergo mutagenesis in different fashions. Signature patterns should be able to distinguish cancer types.

3) The solution should be biological interpretable. Because signatures are not orthogonal, simple regression might lead to solutions that change erratically when small perturbation is made in the observation. Mathematical optimization also might pick the wrong signature. Especially in case of collinearity, LASSO does not provide guarantee to pick the correct predictor. Researchers now solve this problem by simply taking away the majority of predictors they believe to be inactive. SigLASSO allows users to supply domain knowledge to guide the variable selection in a soft manner.

4) The solution should reflect the level of ascertainment. Especially in WXS, low mutation count is often a severe obstacle for assigning signatures due to undersampling. Care should be taken to not overfit the data.

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

ROBUST

2.1 WGS scenario: renal cancer datasets, prior matters

We benchmarked the two methods using 35 Whole-genome sequenced papillary kidney cancer samples (Figure 3, REF). The median mutation count is 4528 (range: 912-9257). We found without prior, both SigLASSO and deconstructSigs showed high contribution from signature 3 and 8, which were thought not active

in pRCC from previous studies and currently there lacks biological support to rationalize them existence in pRCC (REF).

However, if we just “subset” the signatures and take the ones are active from previous studies, the signature profile is completely dominated by signature 5 with only roughly 30-40% mutations assigned with signature, indicating possible underfitting.
When sigLASSO takes into prior knowledge of active signatures, the assignment increases to around 70% in most cases. The backbone signature is signature 5, which is in line with previous reports. SigLASSO also assigned a small portion of mutations to signature 3 and 13.

Fig 3: SigLASSO and deconstructSigs performance on 35 WGS papillary renal cell carcinoma samples. Bars represent the fraction of mutation assigned with signatures. Samples are sorted by the fraction of signature SigLASSO assigned. Pie charts show the total signature contribution when summing up all 35 samples.

2.2 WXS scenario: esophageal carcinoma, our method is sensitive to mutation counts

Then we moved to run the two methods on 181 whole-exome sequenced esophageal carcinoma samples with at least 20 mutations. The median mutation count is 78 (range: 23-1001), which is a low mutation counts situation. No prior is used because COSMIC does not have active signatures in esophageal cancers. SigLASSO only assigns signatures to 20-40% of the mutations. There is a weak but significant positive correlation between mutation count and fraction of mutation with signature inferred (correlation = 0.07, $p<0.001$, Supplement 1). In contrast, deconstructSigs assigns signatures to more than 80% and often 100% of the total mutation. The fractions of signatures assigned have no significant correlation with total mutation counts ($p>0.05$).

Signature 5 (“age”) dominates the solution from SigLASSO, followed by signature 3, 25, 9 and 1 (Fig 4A). In deconstructSigs, the dominating signature is 25, followed by 3, 1, 9 and 24. According to COSMIC, signature 5 and 1 are the aging signature. They are the only two signatures that are active in all cancers shown on COSMIC. We expected age signature to be also active in non-pediatric, esophageal cancers. Meanwhile, the etiology for signature 25 is unknown but only observed in Hodgkin’s lymphomas cell line. Similarly, signature 9 is linked with AID activity in leukemia and lymphoma. We believe these two signature assignments are not biologically interpretable.

Last, we demonstrated SigLASSO could help distinguish different histological types of esophageal cancer (Fig 4B). In the Adenocarcinoma type, SigLASSO found more signature 5 but less signature 3. DeconstructSigs found slightly more signature 3 but less signature 25. ANOVA showed …

Real cancer mutational profiles are likely noisier than our simulation and exhibit highly nonrandom distribution of signatures. They might explain the performance disparity on simulated and read datasets.

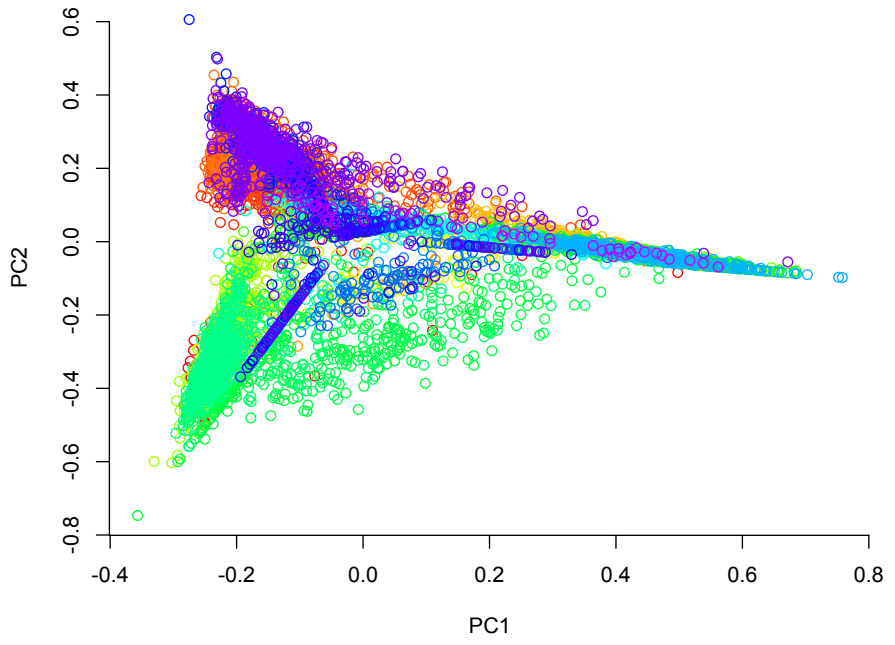
Fig 4: SigLASSO and deconstructSigs performance on 181 WXS esophageal carcinoma samples. **A**: Top two panels: bars represent the fraction of mutation assigned with signatures. Samples are sorted by the fraction of signature SigLASSO assigned. Pie charts show the total signature contribution when summing up all samples. Bottom panel: bars represent the according mutation counts in samples. **B**: Pie charts show the total signature contribution in two different histological subtypes assigned by sigLASSO and deconstructSigs.

2.3 Performance on 8,893 TCGA samples

We ran SigLASSO with step-by-step set-ups and deconstructSigs on 8,893 TCGA tumors (34 cancer types, Supplemental X) that have >20 mutations. The results are shown in figure X.

We noticed, after applying either subsampling or L1 penalty, the results became sparser compared to single regression. Combining both leaded to even higher sparsity. Yet, without giving priors, signature 3 and 25 contributed large portions to the mutations in almost every cancer. Based on previous studies, signature 3 and 25 are believed to be inactive in most cancers. This issue is also observed, to a greater extent, in deconstructSigs. After adding in cancer type-specific priors from large-scale signature studies, sigLASSO results showed significant improvement, with “aging” signature 1 and 5 dominating.

SigLASSO provided better clustering of cancer types based on the signature distribution as shown in the PCA plot (Fig5B? STL2MG: I pasted them below, we should have another figure pack discussion). ANOVA shows the cancer types show distinguishable signature patterns…


PC2
PC1
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8

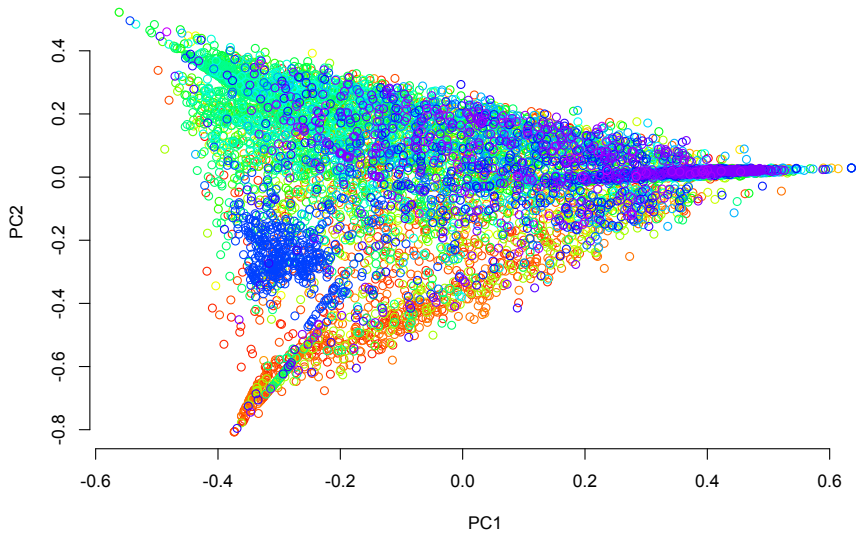

PC2
PC1
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6

Fig5: A heat map of step-by-step SigLASSO performance and deconstructSigs on 34 cancer types.

Discussion

Recently, decomposing cancer mutations into a linear combination of signatures provides invaluable insights in cancers (REF). Though inferring mutational signatures and the latent mutational processes, researchers are able to start better understanding one of the fundamental driving force of cancer initiation and development: mutagenesis.

How to leverage on results from large-scale signature studies and apply to a small set of samples is a very practical problem for many researchers. While this might seem to be a simple regression problem at first, the core question is how to promote sparsity and prevent over- and underfiting. Researchers learned from signature studies in large-scale cancer datasets that mutational signatures are not all active in one sample. In most tumor cases, only a few signatures dominate. A recent signature study summary shows 2-to-13 known signatures are observed in a given cancer type, which might include hundreds and even thousands samples. Moreover, the solution should be aware of data complexity and parameterize accordingly to avoid over- and underfiting. Last, mutational signatures are not orthogonal due to their biological nature. Colinearity of the signatures will lead instable fittings that change erratically with even slight perturbation of the observation.

Shantao 2/2/2018 11:28 PM
Deleted:

DeconstructSigs is the first tool to identify signatures even in a single tumor. Here, we developed SigLASSO, providing a more mathematically rigorous alternate. Unlike deconstructSigs paving a forward selection path, SigLASSO uses L1 to penalize the coefficients for signature selection and promoting

sparsity. By fine-tuning the penalizing terms, SigLASSO is able to further exploit previous signature studies from large cohorts and promote signatures that are believed to be active.

Moreover, under the current model, cancer draws mutations from a multinomial distribution of all active cancer signatures and then further draw from the multinomial nucleotide context distribution given by the signature. The sampling is usually stable with abundant mutations in whole genome sequencing. However, in whole exome sequencing, cancer samples having less than 50 mutations are common. Thoese mutations are first divided into several signatures and then categorized further into 96 types based on the nucleotide composition. With mutation number less then a few hundreds; undersampling becomes a significant obstacle for reliable signature identification.

SigLASSO tries to take a conservative approach and utilizes subsampling to assess the signature inference ascertainment. So that the number of assigned signatures (model complexity) is informed by the data complexity. Likewise, SigLASSO does not specify a noise level explicitly beforehand (in contrast, deconstructSigs specifies a noise level of 0.05 to derive the cut-off of 0.06 for stopping) but uses cross validating to parameterize. In general, SigLASSO let data itself control the model complexity.

Last, due to the colinearity nature of the signatures, pure mathematical optimization might lead to picking wrong signatures that are highly correlated with the true active ones. To overcome this problem, SigLASSO allows researchers to incorporate domain knowledge to guide signature identification. We showcased its performance on real cancer dataset. Although we lack the ground truth of the operative mutational signatures in the tumors, nonetheless we have several reasonable believes about the signature solution. SigLASSO produced signature solutions that are more biologically interpretable, better align with our current

knowledge and believes about mutational signatures and well distinguish cancer types and histological subtypes.