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

[[STL: I marked a few words that I am not sure about in ugly green]]

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. The solution should be sparse and biologically interpretable. Previously published methods use empirical forward selection or iterate signature combinations using 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, LASSO organically 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 by parameterizing using cross-validation and subsampling.

Shantao 1/2/2018 1:12 PM
Formatted: Font:12 pt, Bold

Shantao 1/2/2018 1:12 PM
Formatted: Font:12 pt, Bold, Highlight

Shantao 1/2/2018 1:12 PM
Formatted: Font:12 pt, Bold

Shantao 11/16/2017 8:24 PM
Deleted: break down

Shantao 1/2/2018 12:58 PM
Deleted: all

Shantao 12/11/2017 6:58 PM
Formatted: Font:Italic, Highlight

Shantao 1/2/2018 12:58 PM
Deleted: it

Shantao 12/11/2017 6:37 PM
Deleted: more mathematically justified

Shantao 12/11/2017 6:48 PM
Deleted: problem

Shantao 11/16/2017 5:00 PM
Deleted: current approach of

Shantao 11/16/2017 5:00 PM
Deleted: , leading to a more reliable and interpretable signature solution

Shantao 11/16/2017 5:01 PM
Deleted: our method is automatically parameterized based on cross-validation and subsampling. The

Shantao 11/16/2017 5:01 PM
Deleted: This objective, robust approach promotes data replicability and fair comparison across samples.

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 thirty 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.

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]. 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. By penalizing the L1 norm of coefficients, the algorithm 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 method leaves leeway for noises and unidentified signatures. Last, our method is 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

Shantao 1/2/2018 1:06 PM
Deleted: the

Shantao 1/2/2018 1:10 PM
Deleted: mutation

Shantao 1/2/2018 1:11 PM
Deleted: Understanding

Shantao 1/2/2018 1:12 PM
Formatted: Highlight

Shantao 1/2/2018 1:14 PM
Deleted: empirical

Shantao 1/2/2018 1:15 PM
Formatted: Highlight

Shantao 1/2/2018 1:16 PM
Deleted: LASSO

Shantao 1/2/2018 1:16 PM
Deleted: s

Shantao 11/18/2017 6:51 PM
Deleted: , leading to a more reliable and interpretable signature solution

Shantao 1/2/2018 1:17 PM
Deleted: automatically

Shantao 1/2/2018 1:17 PM
Deleted: objective, robust

Shantao 1/2/2018 1:18 PM
Formatted: Highlight

Shantao 1/2/2018 1:17 PM
Deleted: data

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

Shantao 1/2/2018 1:34 PM
Deleted: from

Shantao 1/2/2018 1:36 PM
Formatted: Font:Italic

Shantao 1/2/2018 1:37 PM
Formatted: Font:Italic

Shantao 1/2/2018 1:37 PM
Formatted: Font:Italic

Shantao 11/18/2017 7:10 PM
Deleted: weighting

Shantao 1/2/2018 1:43 PM
Deleted: Of

Shantao 12/11/2017 7:28 PM
Deleted: (coefficients)

Shantao 12/11/2017 7:28 PM
Deleted: .

Shantao 11/18/2017 7:11 PM
Deleted: both

Shantao 12/11/2017 7:31 PM
Moved (insertion) [2]

Shantao 12/11/2017 7:31 PM
Deleted:

Shantao 12/11/2017 7:31 PM
Deleted:

Shantao 11/18/2017 7:12 PM
Moved down [1]: Mutation count is an important factor for signature identification.

Shantao 11/18/2017 7:12 PM
Deleted: I is a vector of indicator functions of whether a signature should be penalized. If we have strong prior belief that a signature should be active, the corresponding I is zero. The corresponding coefficient for the signature is then not penalized. .

Shantao 12/11/2017 7:31 PM
Moved up [2]: Mathematically, LASSO is equivalent to a Bayesian linear regression framework with Laplace prior.

Shantao 11/18/2017 7:12 PM
Moved (insertion) [1]

Shantao 11/18/2017 7:13 PM
Deleted: for

Shantao 1/2/2018 1:46 PM
Deleted: account

Shantao 1/2/2018 1:47 PM
Deleted: less mutations

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).

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.

Testing 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.

Shantao 1/2/2018 1:47 PM
Deleted: we

Shantao 1/2/2018 1:47 PM
Deleted: we

Shantao 1/2/2018 1:47 PM
Deleted: SigLASSO

Shantao 12/11/2017 7:32 PM
Deleted: have

Shantao 11/18/2017 7:14 PM
Deleted: 5

Shantao 11/19/2017 12:15 AM
Deleted: true

Shantao 1/2/2018 1:53 PM
Deleted: truth

Shantao 1/2/2018 1:53 PM
Deleted: false positive and

Shantao 1/2/2018 1:53 PM
Deleted: We realized the real cancer mutational profiles are much noisier than our simulation and exhibit highly nonrandom distribution of signatures.

Shantao 1/2/2018 1:54 PM
Deleted: a

Shantao 1/2/2018 1:24 PM
Deleted: -

Shantao 1/2/2018 1:54 PM
Deleted: a

Shantao 1/2/2018 1:54 PM
Deleted: analysis

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).

Shantao 12/11/2017 7:34 PM
Deleted: published

Results

Shantao 11/25/2017 11:30 PM
Deleted: .

1. Performance on simulated dataset

Shantao 11/25/2017 11:28 PM
Deleted: .

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.

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.

Shantao 11/25/2017 11:28 PM
Deleted: .

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

Shantao 11/26/2017 2:48 PM
Deleted: LASSO if more computationally efficient than forward selection. . ... [1]

2. Performance on real dataset

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.

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 ($p>0.05$).

Shantao 12/11/2017 10:33 PM
Deleted: .

Shantao 12/11/2017 10:33 PM
Deleted: .

Shantao 12/11/2017 10:39 PM
Formatted: Highlight

Shantao 12/11/2017 10:39 PM
Formatted: Highlight

Shantao 12/11/2017 10:33 PM
Deleted: Priors are from previous large-scale Pan-cancer studies.

Shantao 12/11/2017 10:44 PM
Deleted: .

Shantao 12/11/2017 10:45 PM
Deleted: W

Shantao 12/11/2017 10:45 PM
Deleted: mutation counts >

Shantao 1/2/2018 3:04 PM
Formatted: Highlight

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.

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.

2.3 Implications in infer signature changes in tumor evolution?

[[Showcase first infer active signature from all the mutations/samples…then feed in these *active signatures* to dissolve the early/late mutation set. But the problem is, why is subsetting not good (or even better) for this problem]]

Shantao 12/11/2017 10:59 PM
Deleted:

Shantao 12/11/2017 11:00 PM
Formatted: Font:Italic

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. DeconstructSigs is the first tool to identify signatures even in a single tumor. Here, we developed SigLASSO, providing a more mathematically rigorous alternate.

Shantao 12/11/2017 11:31 PM
Formatted: Highlight

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.

Shantao 12/11/2017 11:52 PM
Deleted: .

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

Shantao 1/2/2018 4:09 PM
Formatted: Highlight

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.**

[[Signature similarity and correlation between signatures?]]