[[STL2MG: uncertain words/phrases highlighted in ugly green]]

Shantao 3/11/2018 2:21 PM
Formatted: Highlight

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

Shantao 3/11/2018 2:48 PM
Deleted: regression

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. Although the underlying true mutational distributions are often impossible to obtain, researchers have certain sensible/plausible beliefs/ideas/notions about the assignment. For instance, existing 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 formulate the problem as a LASSO linear regression and accordingly develop 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 harmoniously 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 3/11/2018 2:21 PM
Formatted: Highlight

Shantao 3/3/2018 9:40 PM
Formatted: Highlight

Shantao 3/4/2018 8:33 PM
Deleted: Past

Shantao 3/9/2018 3:03 PM
Deleted: alternatively

Shantao 3/12/2018 3:49 PM
Deleted: SigLASSO

Shantao 3/12/2018 3:49 PM
Deleted: SigLASSO

Shantao 3/11/2018 2:31 PM
Formatted: Highlight

Shantao 3/4/2018 8:44 PM
Formatted: Highlight

Shantao 3/9/2018 3:04 PM
Deleted:

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 [REF]. Multiple endogenous and exogenous mutational processes drive cancer mutagenesis and leave distinct fingerprints [REF]. 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 happened 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.

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 the latent mutational processes 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. Likewise, a very specific AID mutation process, biologically involved in antibody diversification, should not be observed outside B

Shantao 3/11/2018 2:25 PM
Deleted: signatures

cell lymphoma. Besides, we are also motivated by the Occam's principle here. A sparser solution is preferred as it explains the observation in a simpler fashion.

Previously published methods use forward selection with an empirically derived stopping criterion or iterate all combinations by brutal force (REF). There are also approaches using linear programming (REF), which is not efficient regarding optimization. Here, we formulate the task as a more mathematically rigorous LASSO linear regression problem. Out approach is the first one that explicitly penalizes the model complexity by regularization. We use L1 norm as the regularizer as L0 norm (cardinality of active signatures) is designed but cannot be effectively optimized. On the other hand, L2 norm leads to many small, non-zero coefficients, which are hard to interpret biologically. By penalizing the L1 norm of coefficients, the algorithm is efficient and produces sparse, biologically interpretable solutions. Additionally, this approach is able to harmoniously integrate biological prior knowledge into the solution by fine-tuning penalties on the coefficients. Compared with current approach of hardly subseting signatures before fitting, our method leaves leeway for noises and unidentified signatures. Last, unlike previous methods, sigLASSO is aware of data complexity such as mutational number and patterns in the observation. Our method is automatically parameterized based on cross-validation and subsampling, allowing data complexity to inform model complexity. This prudent approach promotes results replicability and fair comparison across datasets.

Shantao 3/11/2018 2:56 PM
Comment [1]: Should we bring up the previous methods (e.g. deconstrucSigs) are all pursuing sparsity too, but in more or less implicit ways.

Shantao 3/9/2018 11:13 PM
Deleted:

Shantao 3/11/2018 2:27 PM
Deleted: (

Shantao 3/11/2018 2:27 PM
Deleted:)

Shantao 3/12/2018 6:47 PM
Deleted:

Shantao 3/11/2018 2:27 PM
Deleted: alternatively

Shantao 3/11/2018 2:27 PM
Deleted: it

Shantao 3/11/2018 2:28 PM
Deleted: in optimization

Shantao 3/11/2018 2:29 PM
Deleted: L2 norm, o

Shantao 3/11/2018 2:30 PM
Deleted: and

Shantao 3/11/2018 2:31 PM
Deleted: organically

Shantao 3/11/2018 2:32 PM
Deleted: soft prior

Shantao 3/12/2018 3:49 PM
Deleted: SigLASSO

Shantao 3/11/2018 2:34 PM
Formatted: Highlight

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 from new samples into a linear combination of signatures. Mathematically, the problem is formulated as the following nonnegative regression problem:

$$\min_{W \in R^+} \|M - WS\|_2^2$$

The mutation matrix, *M*, contains mutations of each sample cataloged into 96 trinucleotide 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). LASSO is equivalent to a Bayesian linear regression framework with Laplace prior.

$$\min_{W \in R^+} (\|M - WS\|_2^2 + \sum \lambda I \|W\|)$$

λ is parameterized by 12-fold cross validation. Cross validation was done by splitting 96 trinucleotide contexts into 12 (a divisor of 96) groups and rotationally holding off one group (8 trinucleotide contexts) for testing and train the model on the rest 11 groups (the rest 88 trinucleotide contexts). A correct signature solution should be inferable by using ~92% (11/12) of the trinucleotide contextual information and predict well the rest 8%. Any over- or underfitting will lead to higher error in predication on the test set. We use the largest λ (which leads to a sparser solution) that gives mean square error (MSE) within 3 standard deviances (SD) of the minimum. *I* is an indicator vector, indicating whether a certain signature should be fully penalized (i.e. 1), only partially penalized (e. g. 0.5) or not at all (i.e. 0). It should be tuned to reflect the level of confidence in prior knowledge.

Shantao 3/11/2018 6:11 PM
Deleted: observed in

Shantao 3/11/2018 2:16 PM

Shantao 3/11/2018 2:16 PM
Deleted: 2

Shantao 3/11/2018 6:14 PM
Deleted: broken down

Shantao 3/12/2018 3:49 PM
Deleted: SigLASSO

Shantao 3/11/2018 6:18 PM
Deleted: Mathematically,

Shantao 3/11/2018 2:16 PM
Deleted: . ... [1]

Shantao 3/4/2018 8:46 PM
Deleted: 10

Shantao 3/11/2018 6:20 PM
Deleted: smallest

Shantao 3/12/2018 2:08 PM
Formatted: Font:Italic, Highlight

Shantao 3/12/2018 2:08 PM
Formatted: Highlight

Mutation count is a major factor affecting signature identification. To assess the solution stability and prudently adjust for lower signature ascertainment when fewer mutations are observed, sigLASSO performs subsampling. At each subsampling step, it samples 50% mutations, solves the LASSO 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).

Shantao 3/11/2018 6:49 PM
Deleted: an important

Shantao 3/12/2018 3:49 PM
Deleted: SigLASSO

Shantao 3/11/2018 6:49 PM
Deleted: regression

Shantao 3/12/2018 3:47 PM
Deleted: SigLASSO

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.

Shantao 3/12/2018 3:47 PM
Deleted: SigLASSO

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) and sigLASSO without prior knowledge of the underlying signature. To evaluate the performances, we compared the inferred signature distribution with the simulated distribution and calculated 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). Prior knowledge on active signatures in various cancer types was also extracted from this source.

SigLASSO software suite

SigLASSO accepts processed mutational spectrums. We provide simple script to help parse mutational spectrums from VCF files. SigLASSO allows the users to specify biological priors (i.e. signatures that should be active or inactive), subsampling steps and subsampling cutoff. SigLASSO uses the 30 COSMIC signatures by default. Users are also given the option to supply customized signature files. LASSO is computationally efficient. Using default settings, the program could successfully decompose a cancer WGS sample data in less than a minute 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.

Shantao 3/11/2018 6:53 PM
Deleted: s used in SigLASSO

Shantao 3/11/2018 6:54 PM
Deleted: ere

Shantao 3/11/2018 6:54 PM
Deleted: (vcf files or)

Shantao 3/11/2018 6:55 PM
Deleted: It

Shantao 3/11/2018 6:56 PM
Deleted: also

Shantao 3/11/2018 6:56 PM
Deleted: a few seconds

Shantao 3/12/2018 3:48 PM
Deleted: SigLASSO

Shantao 3/12/2018 3:48 PM
Deleted: SigLASSO

Even in case of a program that recovers all signatures perfectly but also oblivious about the noise, its MSE will be the square of noise level, which is 0.01 in this case. Likewise, its 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 (0.2 and 0.1 respectively) and high/low mutation counts (5,000 and 200 respectively). Error bars indicate one standard deviation (SD) of ten repeats.

When mutation number decreases, there is an increase of uncertainty in sampling, which is negligible in high mutation scenarios. As expected, the MSE jumped into the 0.1-0.3 range for both low and high noise setups. 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 typically cannot be obtained. Previous large-scale signature studies largely rely on mutagen exposure association from patient records and biochemistry knowledge on mutagenesis. Here, we illustrated the outputs of different models and compared the results with existing signature knowledge. Although there is no golden standard to evaluate the performance, we do have a few reasonable expectations about the solution.

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

Shantao 3/12/2018 12:13 PM
Deleted: we introduce

Shantao 3/12/2018 12:14 PM
Deleted: number cases

Shantao 3/12/2018 12:14 PM
Deleted: to

Shantao 3/12/2018 1:48 PM
Deleted: simply

Shantao 3/12/2018 1:48 PM
Deleted: .

Shantao 3/12/2018 1:49 PM
Deleted: in a given cancer sample or type

signature should not be expected in tissues unexposed. Previous signature studies suggest a sparse distribution of signatures among cancer samples and types. Existing signature identifying methods try to implicitly achieve sparse solutions by forward selection or pre-select the signature set for fitting.

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

3) Robustness: Solutions should be robust and reproducible. Signatures are not orthogonal, thus simple regression might lead to solutions that change erratically when small perturbation is made in the observation. Moreover, 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 avoid overfitting the data.

4) 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. Even 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.

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

2.1 WGS scenario: renal cancer datasets, prior knowledge matters

We benchmarked the two methods using 35 Whole-genome sequenced papillary kidney cancer samples (Figure 3, REF). The median mutation count is 4,528 (range: 912-9,257). We found without prior, both sigLASSO and deconstructSigs showed high contribution from signature 3 and 8, which were found inactive in

Shantao 3/12/2018 2:[illegible] PM
Deleted: observed

Shantao 3/12/2018 1:54 PM
Deleted: different

Shantao 3/4/2018 9:19 PM
Deleted: 3

Shantao 3/4/2018 9:20 PM
Deleted: signatures are not orthogonal, simple regression might lead to solutions that change erratically when small perturbation is made in the observation. M

Shantao 3/4/2018 9:20 PM
Deleted: also

Shantao 3/4/2018 9:21 PM
Deleted: Especially in case of collinearity,

Shantao 3/4/2018 9:19 PM
Deleted:

Shantao 3/4/2018 9:18 PM
Deleted: 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.

Shantao 3/12/2018 2:01 PM
Deleted: find

Shantao 3/12/2018 3:48 PM
Deleted: SigLASSO

Shantao 3/12/2018 2:09 PM
Deleted: though

Shantao 3/12/2018 2:09 PM
Deleted: t

Shantao 3/12/2018 2:09 PM
Deleted: not

pRCC from previous studies and currently there is no biological support to rationalize them existence in pRCC (REF).

However, if we naively "subset" the signatures and take the ones are found 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: 181 esophageal carcinoma

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: 23-1,001), 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. In contrast, deconstructSigs assigns signatures to more than 80% and often 100% of the total mutation.

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,

Shantao 3/12/2018 2:09 PM
Deleted: lacks

Shantao 3/12/2018 2:10 PM
Deleted: just

Shantao 3/12/2018 3:27 PM
Deleted: . ... [2]

Shantao 3/12/2018 3:48 PM
Deleted: SigLASSO

Shantao 3/12/2018 8:04 PM
Deleted: , our method is sensitive to mutation counts

Shantao 3/12/2018 8:04 PM
Deleted: 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).

Shantao 3/12/2018 8:04 PM
Deleted: The fractions of signatures assigned have no significant correlation with total mutation counts (p>0.05).

Shantao 3/12/2018 3:48 PM
Deleted: SigLASSO

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 and likely caused by noise or yet unknown signatures.

Shantao 3/12/2018 6:44 PM
Deleted: .

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.

Shantao 3/12/2018 3:48 PM
Deleted: SigLASSO

Shantao 3/12/2018 3:48 PM
Deleted: SigLASSO

Shantao 3/12/2018 6:44 PM
Deleted: 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.

Shantao 3/12/2018 3:48 PM
Deleted: SigLASSO

2.3 Performance on 8,892 TCGA samples

Shantao 3/12/2018 8:43 PM
Deleted: 3

We ran sigLASSO with step-by-step set-ups and deconstructSigs on 8,892 TCGA tumors (31 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.

Then we moved on and tested how signature identified could discern homology repair (HR) defect samples. We pulled 229 samples with putatively loss of function of BRCA1/2 mutations in 25 cancer types. Then we scrutinized the signature distribution on samples with mutant BRCA 1/2 and samples with matched cancer types.

Fig5: A heat map of step-by-step sigLASSO performance and deconstructSigs on 33 cancer types.

2.4 Robustness assessment on down sampled profiles

Targeted sequencing, low sequencing depth and certain cancer types could all lead to low mutation counts. To assess the performance of sigLASSO under low mutation counts, we performed down sampling on 30 pRCC WGS samples that have more than 3,000 mutations. Each down sample size was repeated ten times.

Shantao 3/12/2018 3:48 PM
Deleted: SigLASSO

Shantao 3/12/2018 8:43 PM
Deleted: 3

Shantao 3/12/2018 8:30 PM
Deleted: 4

Shantao 3/12/2018 8:13 PM
Deleted: SigLASSO provided better clustering of cancer types based on the signature distribution as shown in the PCA plot (Fig5B ANOVA shows the cancer types show distinguishable signature patterns…

Shantao 3/12/2018 3:48 PM
Deleted: SigLASSO

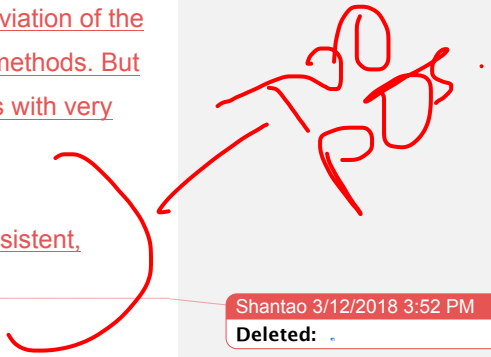
Shantao 3/12/2018 10:27 PM
Deleted: 4

Shantao 3/12/2018 3:36 PM
Deleted:

Shantao 3/12/2018 3:55 PM
Formatted: Font:Bold

We noticed sigLASSO assigned more mutations with signatures when the mutation counts increases from extremely low (Fig 6). It stabilized after ~100 mutations. In contrast, deconstructSigs assigned fewer mutations with signatures as the mutation number increases. As expected, mean standard deviation of the assignments decreases as the mutations count increases for both methods. But sigLASSO exhibits a significantly lower fluctuation. Even in samples with very few mutations, the deviation is small.

In conclusion, sigLASSO is resilient to low mutation counts and consistent, producing robust results.



Shantao 3/12/2018 3:52 PM
Deleted:

Fig6:

Discussion

Recently, decomposing cancer mutations into a linear combination of signatures provides invaluable insights into cancer biology (REF). Through inferring mutational signatures and the latent mutational processes, researchers gained better understanding one of the fundamental driving force of cancer initiation and development: mutagenesis.

Shantao 3/11/2018 2:35 PM
Deleted: s

Shantao 3/11/2018 2:35 PM
Deleted: are able to start

How to leverage on results from large-scale signature studies and apply to a small set of incoming samples is a very practical problem for many researchers. While this might seem to be a simple linear system 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 or cancer type. In most tumor cases, only a few signatures prevail. A recent signature summary shows 2-to-13 known signatures are observed in a given cancer type [REF], which might include hundreds and even thousands samples. Not only sparse solutions are

Shantao 3/11/2018 2:36 PM
Deleted: a

Shantao 3/11/2018 2:36 PM
Deleted: regression

Shantao 3/11/2018 2:38 PM
Deleted: dominate

Shantao 3/11/2018 2:39 PM
Deleted: study

both biologically sound and better interpretable, but also are motivated by the Occam's razor principle, which prefers the simplest solution that explains the observation.

Moreover, the designed method should be aware of data complexity and parameterized 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.

DeconstructSigs is the first tool to identify signatures even in a single tumor. It archives sparsity using a stopping criterion for adding in new signatures in forward selection. Here, we developed and presented sigLASSO, providing a more mathematically rigorous alternate. Unlike deconstructSigs paving a forward selection path and fitting an unconstrained linear model at every step, sigLASSO uses L1 norm to penalize the coefficients for signature selection, thus promote 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.

Additionally, as sequencing cost drops rapidly, we expect to see more cancer samples getting whole genome sequenced. The vast amount of cancer genomics data will discern more occult 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 context). Traditional simple solver would give infinitude (noiseless) or unstable (noisy) solutions in this underdetermined linear system. However, by assuming the solution is sparse, we are able to apply regulation to achieve a simpler, sparse solution (basic pursuit/basic pursuit denoising).

Moreover, under the current generative model, cancer draws mutations from a multinomial distribution of all active cancer signatures and then further draw from

Shantao 3/11/2018 2:40 PM
Deleted: solution

Shantao 3/12/2018 3:48 PM
Deleted: SigLASSO

Shantao 3/12/2018 3:48 PM
Deleted: SigLASSO

Shantao 3/11/2018 2:41 PM
Deleted: and

Shantao 3/11/2018 2:41 PM
Deleted: ing

Shantao 3/12/2018 3:48 PM
Deleted: SigLASSO

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

SigLASSO tries to take a prudent 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. This knowledge input could be cancer-type specific signatures, patient clinical information (e.g. smoking history, chemotherapy etc.) and many others. We showcased its performance on real cancer dataset. Although we lack the ground truth of the operative mutational signatures in 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 about mutational signatures and well distinguish cancer types and histological subtypes.

SigLASSO exploits constraints in signature identifying and provides a robust framework to achieve biologically sound solutions. Due to the highly

Shantao 3/11/2018 2:51 PM
Deleted: 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

Shantao 3/11/2018 2:51 PM
Deleted: m

Shantao 3/11/2018 2:51 PM
Deleted: conservative

Shantao 3/12/2018 3:48 PM
Deleted: SigLASSO

Shantao 3/12/2018 3:48 PM
Deleted: SigLASSO

Shantao 3/12/2018 3:48 PM
Deleted: SigLASSO

Shantao 3/11/2018 2:53 PM
Deleted: the

Shantao 3/11/2018 2:53 PM
Deleted: and believes

interdisciplinary nature, identifying signatures in cancer samples is a challenging task. For instance, the confidence level of the prior knowledge should be used to inform the optimum penalties for likely active signatures. Yet right now, it is often arduous, if even possible, to quantify. Nonetheless, sigLASSO offers a framework that empowers researchers to use and integrate their biological knowledge and expertise into the model.