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

Abstract: Multiple mutational processes fuel carcinogenesis. These processes leave characteristic signatures in cancer genomes. Deciphering the signatures of mutational processes operative in cancer can help elucidate the mechanisms underlying cancer initiation and development. This process involves decomposing cancer mutations by nucleotide context into a linear combination of mutational signatures. Previously published methods use empirical forward selection or iterate signature combinations by brute force. However, these methods have inherent limitations that make them difficult to interpret biologically. Here, we developed a software tool, sigLASSO, that formulates the problem as a LASSO linear regression. By parsimoniously assigning signatures to cancer genome mutation profiles, the solution becomes sparse and more biologically interpretable. Additionally, sigLASSO integrates prior biological knowledge harmoniously into the solution by fine-tuning penalties on coefficients. Compared with subsetting signatures before fitting, our method leaves leeway for noise and unknown signatures. The model complexity is informed by the size and complexity of the data through parameterizing using cross-validation and subsampling. In sum, sigLASSO offers a framework that empowers researchers to use and integrate their biological knowledge and expertise into the model.

Introduction

Mutagenesis is a fundamental process underlying cancer development. Examples include spontaneous deamination of cytosines, the formation of pyrimidine dimers by ultraviolet (UV) light, and the crosslinking of guanines by alkylating agents [REF]. Multiple endogenous and exogenous mutational processes drive cancer mutagenesis and leave distinct fingerprints [REF]. Notably, these processes have characteristic mutational nucleotide context biases. Mutation profiling of cancer samples at manifestation has revealed that mutations accumulate over a lifetime; this includes somatic alterations that occur

Shantao 3/22/2018 4:02 PM
Deleted: delineating

SEN HEAVY 3/26/2018 6:04 PM
Comment [1]: Correct? I felt you should set up the gap in the field.

SEN HEAVY 3/26/2018 6:08 PM
Deleted: Although the underlying true mutational distributions are often impossible to obtain, researchers have certaincan make logical assumptions sensible/plausible beliefs/ideas/notions about the assignment. For instance, existing previous mutational signature studies suggest that the solution should be sparse to be biologically interpretable. Previously published methods use empirical forward selection or iterate signature combinations by bruteal force. Here, we formulated the problem as a LASSO linear regression and accordingly developed a software tool, sigLASSO.

SEN HEAVY 3/26/2018 6:08 PM
Deleted: Last

SEN HEAVY 3/26/2018 6:09 PM
Deleted: , the

SEN HEAVY 3/26/2018 6:10 PM
Comment [2]: I would end with a conclusion sentence with broader impact. Feel free to adjust this.

SEN HEAVY 3/26/2018 6:10 PM
Deleted: the

Shantao 3/31/2018 9:50 PM
Formatted: Highlight

SEN HEAVY 3/26/2018 1:39 PM
Deleted: found

SEN 3/26/2018 1:20 PM
Deleted: all

SEN 3/26/2018 1:20 PM
Deleted: the

Shantao 3/31/2018 9:49 PM
Formatted: Highlight

both before cancer initiation and during cancer development. In a generative model, multiple latent processes generate mutations over time, drawing from their corresponding nucleotide context distributions ("mutation signature"). 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 identified at least 30 mutational processes [REF]. Many processes have been recognized and linked with known etiologies, such as aging, smoking, or ApoBEC activity. Investigating the fundamental processes underlying mutagenesis can help elucidate cancer initiation and development.

One major task in cancer research is to leverage signature studies on large-scale cancer cohorts and efficiently assign active signatures to new cancer samples [REF]. Although we do not fully know the latent mutational processes in cancer samples, we can make reasonable and logical assumptions about the solutions of such studies. Here, we aimed to design a computational framework that could meet these expectations. For example, we believe a solution should be sparse as past studies indicate that not all signatures can be active in a single sample or even a given cancer type. An apparent example is, we should not observe UV-associated signatures in tissues that are not exposed to UV. Likewise, we only expect to observe activation-induced cytidine deaminase (AID) mutational processes, which are biologically involved in antibody diversification, in B cell lymphomas. We also prefer a sparser solution as it explains an observation in a simpler fashion, consistent with Occam's principle.

Previously published methods use forward selection with an empirically derived stopping criterion or iterate all combinations by brute force (REF). Other approaches use linear programming (REF), which is not efficient in optimization. Here, we formulated the task as a mathematically rigorous LASSO linear

SEN 3/26/2018 1:29 PM
Deleted: scientists

Shantao 3/31/2018 9:59 PM
Formatted: Highlight

SEN HEAVY 3/26/2018 1:43 PM
Deleted: have the ground truth of

SEN 3/26/2018 1:29 PM
Deleted: they

SEN HEAVY 3/26/2018 1:43 PM
Deleted: do have

SEN HEAVY 3/26/2018 6:11 PM
Deleted: some

SEN HEAVY 3/26/2018 1:43 PM
Deleted: expectations

SEN HEAVY 3/26/2018 7:44 PM
Comment [3]: Solution was vague here. What exactly do you mean?

Shantao 3/31/2018 9:59 PM
Formatted: Highlight

Shantao 3/31/2018 9:59 PM
Formatted: Highlight

SEN HEAVY 3/26/2018 7:46 PM
Comment [4]: I still don't love the word solution here.

Shantao 3/26/2018 7:55 AM
Deleted: regarding

Shantao 3/22/2018 8:04 PM
Deleted: more

regression problem. Our sigLASSO approach penalizes the model complexity by regularization. The most straightforward way to do this would be to use the L0 norm (cardinality of active signatures), but this approach cannot be effectively optimized. Conversely, using the L2 norm flattened out at small values leads to many tiny, non-zero coefficients, which are hard to interpret biologically. Our sigLASSO uses L1 norm, which also promotes sparsity. Meanwhile, L1 norm is a convex map, thus allows efficient optimization. Additionally, this approach is able to harmoniously integrate prior biological knowledge into the solution by fine-tuning penalties on the coefficients. Compared with the current approach of hardly subsetting signatures before fitting, our soft thresholding method leaves leeway for noise and unidentified signatures. Finally, 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 approach promotes result replicability and fair comparison of datasets.

Shantao 3/22/2018 8:19 PM
Deleted: is the first one that explicitly

Shantao 3/26/2018 7:53 AM
Deleted: We use L1 norm as the regularizer as

Shantao 3/26/2018 7:53 AM
Deleted: is designed

Shantao 3/26/2018 7:54 AM
Deleted: small

Shantao 3/31/2018 10:11 PM
Deleted: By penalizing the L1 norm of coefficients, the sigLASSO algorithm is efficient and produces sparse, biologically interpretable solutions.

Shantao 3/31/2018 10:14 PM
Deleted: Last

Shantao 3/22/2018 4:12 PM
Deleted: prudent

Shantao 3/26/2018 9:07 AM
Deleted: across

Material and Methods

Signature identification problem

Mutational processes leave mutations in the genome with distinct nucleotide contexts. Specifically, we considered the mutant nucleotide context and looked one nucleotide ahead and behind. This divides mutations into 96 trinucleotide contexts. Each mutational process carries a unique signature, which is represented by a mutational trinucleotide context distribution (Fig. 1A). Thirty signatures were identified by NMF and clustering from large-scale pan-cancer analysis (REF). Here, our objective was to leverage the pan-cancer analysis and decompose mutations from new samples into a linear combination of signatures. Mathematically, the problem is formulated as the following non-negative regression problem:

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

Shantao 3/26/2018 9:07 AM
Deleted:

Shantao 3/26/2018 9:07 AM
Deleted: on

Shantao 3/26/2018 9:07 AM
Deleted:

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.

$$\min_{W \in R^+} (\|M - SW\|_2^2 + \lambda I \|W\|_1)$$

λ is parameterized by 12-fold (since 12 is a divisor of 96) cross-validation. At every step, we hold off eight trinucleotide contexts as the testing dataset and only fit the signatures on the remaining 88 trinucleotide contexts. We used the largest λ (which leads to a sparser solution) that gives mean square error (MSE) within three standard deviations (SD) of the minimum. *I* is a vector of 30 indicators, each indicating whether a certain signature should be fully penalized (i.e., 1), partially penalized (e.g., 0.5), or not penalized (i.e., 0). This value should be tuned to reflect the level of confidence in prior knowledge.

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 non-negative coefficients) signatures. In the end, signatures are retained that are active in more than τ fraction of all subsampling trials. τ can be set empirically between 0.6 and 0.9 (REF). In our study, we used 0.6 and set subsampling to 100 times unless otherwise specified.

SEN HEAVY 3/26/2018 7:50 PM
Comment [5]: Can you include a brief explanation of the equation in this sentence?

Shantao 3/31/2018 11:36 PM
Deleted: W

Shantao 3/31/2018 11:39 PM
Deleted: $\sum)\lambda I\|W\|)$

Shantao 3/31/2018 10:23 PM
Deleted: used

Shantao 3/26/2018 8:42 AM
Deleted: 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.

Shantao 3/31/2018 10:25 PM
Deleted: ; this approach leads to a sparser solution

Shantao 3/26/2018 8:47 AM
Deleted: n indicator

With the L1 norm to penalization, LASSO will shrink coefficients to zero. In order to get a less biased estimation of coefficients. We performed post-selection inference, following a procedure similar to relaxed LASSO (REF). First, LASSO is equivalent to optimize a constrained linear regression problem:

$$\min_{W \in R^+} (\|M - SW\|_2^2) \quad subject\ to\ \|W\|_1 \leq t$$

Because we restricted $W \in R^+$, L1 norm of W is the sum of all coefficients. In our case, the sum should not exist 1. Therefore, in this step, we ran LASSO on the active signature set selected in previous steps. We used binary search to find a solution that gives coefficients that sum in the range of (0.99, 1]. In case of only one signature in the active set, the procedure degenerates to an ordinary linear regression with coefficient higher than 1 casted to 1. In comparison, deconstructSigs fits coefficients without constrain and then normalizes to one.

Shantao 3/31/2018 11:36 PM
Formatted: Centered

Shantao 3/31/2018 11:47 PM
Formatted: Font:Italic

A schematic illustration of the sigLASSO workflow is shown in Fig. 1B.

Data simulation and model evaluation

First, we downloaded 30 previously identified signatures (http://cancer.sanger.ac.uk/cosmic/signatures, REF). We created a simulated dataset by randomly and uniformly drawing two to eight signatures and corresponding weights (minimum: 0.02). The noise was simulated at various levels with a uniform distribution on 96 trinucleotide contexts. Then, we summed all the signatures and noise to form a mutation distribution. We randomly drew mutations from this distribution with different mutation counts.

Shantao 3/26/2018 9:06 AM
Deleted: N

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 MSE. We also measured the number of false positive and false negative signatures in the solution.

Illustrating on real datasets

To assess the performance of our method on real-world cancer datasets, we used somatic mutations from various cancer types from The Cancer Genome Atlas (TCGA). We downloaded VCF 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 Appendix X.

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

SigLASSO software suite

SigLASSO accepts processed mutational spectrums. We provided simple scripts to help parse mutational spectrums from VCF files. SigLASSO allows users to specify biological priors (i.e., signatures that should be active or inactive), subsampling steps, and the subsampling cutoff. SigLASSO uses 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 can successfully decompose a whole genome sequenced (WGS) cancer sample in less than a minute on a regular laptop (3 GHz i7 CPU, 16 GB DDR3 memory).

We have released SigLASSO as an R package. The updated code is also available on GitHub (https://github.com/ShantaoL/SigLASSO).

Results

1. Performance on a simulated dataset

We first evaluated SigLASSO on a simulated dataset. Both sigLASSO and deconstructSigs performed better with higher mutation number and lower noise (Fig. 2). For a program that recovers all signatures perfectly but is oblivious to

Shantao 3/31/2018 10:28 PM
Deleted: a

Shantao 3/31/2018 10:28 PM
Deleted: the

Shantao 3/31/2018 10:30 PM
Deleted: sequencing

Shantao 3/31/2018 10:30 PM
Deleted: dataset

SEN 3/26/2018 6:32 PM
Comment [6]: I'm not sure where you are submitting, but most journals do not number the results subheadings. Be sure to check journal guidelines.

SEN HEAVY 3/26/2018 6:41 PM
Comment [7]: I feel you need a lead in sentence here. Feel free to adjust

noise, we would expect the MSE to be the square of the noise level. Indeed, the MSE was generally below 0.02 with high mutations and low noise (0.1) in both programs.

A decrease in mutation number leads to an increase of uncertainty in sampling, which is negligible in the high mutation scenarios. As expected, the MSE jumped to the 0.1-0.3 range for both low and high noise setups when the mutation number was low. Thus, the error is dominated by undersampling rather than embedded noise.

Shantao 3/26/2018 9:09 AM
Deleted: . ... [1]

2. Performance on real datasets

We next moved from synthetic datasets to real cancer mutational profiles. One of the limitations of cancer signature research is that the ground truth of real samples typically cannot be obtained. Previous large-scale signature studies largely relied on mutagen exposure 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 no gold standard exists to evaluate the performance, we do have a few reasonable expectations about the solution:

1) Sparsity: One or more signature should be active in a given cancer sample and type. However, not all signatures should be active. Mutational processes are discrete in nature and tied with certain endogenous and environmental factors. An obvious example is that the UV signature should not exist in unexposed tissues. Previous signature studies suggest a sparse distribution of signatures among cancer samples and types. Existing signature identifying methods aim to implicitly achieve sparse solutions by forward selection or pre-selection of 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 between cancer types. It is unrealistic to have the same or similar distribution of signatures in all cancer types, as they have divergent endogenous biological features and environmental exposures.

3) 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 whole exome sequencing (WES), low mutation count is often a severe obstacle for assigning signatures due to undersampling. Care should be taken to avoid overfitting.
4) Biological interpretability: The solution should be biological interpretable. Because of the biological nature of co-linearity in the signatures, simple mathematical optimization might pick the wrong signature. Even LASSO does not provide a guarantee to pick the correct predictor. Researchers now solve this problem by simply removing the majority of predictors they believe to be inactive. SigLASSO allows users to supply domain knowledge to guide the variable selection in a soft thresholding manner.

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

2.1 WGS scenario using renal cancer datasets

We benchmarked the two methods using 35 WGS papillary kidney cancer samples (Fig. 3, REF). The median mutation count was 4,528 (range: 912-9,257). We found that without prior knowledge, both sigLASSO and deconstructSigs showed high contributions from signature 3 and 8; these signatures were found to be inactive in papillary renal cell carcinoma (pRCC) in

previous studies and currently no biological support rationalizes their existence in pRCC (REF).

However, if we naively "subset" the signatures and take the ones that were found to be active in previous studies, the signature profile is completely dominated by signature 5, to which only roughly 30-40% mutations are assigned with signature. This finding suggests possible underfitting.

Shantao 3/31/2018 11:03 PM
Deleted: of the data

When sigLASSO took into account the prior knowledge of active signatures, the assignment increased to around 70% in most cases. The backbone signature was signature 5, which is in line with previous reports. SigLASSO also assigned a small portion of mutations to signature 3 and 13.

2.2 WES scenario using esophageal carcinoma datasets

We next aimed to evaluate the two methods on 181 WES esophageal carcinoma samples with at least 20 mutations. The median mutation count was 78 (range: 23-1,001), which is considerably lower than WGS but typical for WES. We did not use any prior knowledge because COSMIC does not have active signatures in esophageal cancers.

SigLASSO only assigned signatures to 20-40% of the mutations. In contrast, deconstructSigs assigned signatures to over 80% and often 100% of the total mutations.

Shantao 3/31/2018 11:06 PM
Comment [8]: Not too sure after introducing the "relaxed LASSO" thing

Signature 5 ("age") dominated the solution from sigLASSO, followed by signature 3, 25, 9, and 1 (Fig. 4A). According to COSMIC, signature 5 and 1 are the aging signatures and are the only two signatures that are active in all cancers. Thus, it is logical that these two aging signatures would be also active in non-pediatric, esophageal cancers. In deconstructSigs, the dominating signature was 25,

followed by 3, 1, 9, and 24. The etiology for signature 25 is unknown and is only observed in a Hodgkin's lymphoma 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 are likely caused by noise or yet unknown signatures.

Next, we demonstrated that sigLASSO could help distinguish between different histological types of esophageal cancer (Fig. 4B). In the adenocarcinoma type, sigLASSO found more of signature 5 and less of signature 3. DeconstructSigs found slightly more of signature 3 and less of signature 25.

Real cancer mutational profiles are likely noisier than our simulation and exhibit highly non-random distribution of signatures. This might explain the performance disparity between the simulated and real datasets.

Shantao 3/26/2018 9:02 AM
Deleted: on

Shantao 3/26/2018 9:02 AM
Deleted: d

2.3 Performance on 8,892 TCGA samples

We ran sigLASSO with step-by-step set-ups and deconstructSigs on 8,892 TCGA tumors (31 cancer types, Supplemental X) with >20 mutations. The results are shown in Figure X.

We noticed that after applying either subsampling or L1 penalty the results became sparser compared to a single regression. Combining both subsampling and L1 penalty led to an 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. We observed this issue to an even greater extent in deconstructSigs. After adding in cancer type-specific priors from large-scale signature studies, sigLASSO results showed significant improvement, with the "aging" signatures 1 and 5 dominating.

2.4 Robustness assessment on downsampled profiles

Targeted sequencing, low sequencing depth, and certain cancer types can all lead to low mutation counts. To assess the performance of sigLASSO under low mutation situations, we performed downsampling on 30 pRCC WGS samples with over 3,000 mutations. We repeatedly downsampled every sample at each size ten times.

We found that sigLASSO assigned more mutations with signatures when the mutation count increased from extremely low (Fig. 6) and stabilized after ~100 mutations. In contrast, deconstructSigs assigned fewer mutations with signatures as the mutation number increased. deconstructSigs removes signatures with less than 6% relative weights after the coefficient inference. Therefore, we think this decrease of assignment is because deconstructSigs discovered more signatures with small weights as mutation number went up. As expected, the mean standard deviation of the assignments decreased as the mutation count increased for both methods. However, sigLASSO exhibited a significantly lower fluctuation. Even in samples with very few mutations, the deviation was small.

Thus, in low mutation count situations, sigLASSO is able to produce consistent and robust results.

We next tested how the signatures identified could discern homology repair (HR) defect samples. We pulled 229 samples with putative loss-of-function 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.

Shantao 3/26/2018 8:03 AM
Deleted: resilient to low mutation

Shantao 3/26/2018 8:04 AM
Deleted: counts and consistent,

Shantao 3/26/2018 8:04 AM
Deleted: ing

SEN HEAVY 3/26/2018 7:19 PM
Comment [9]: This seems out of place. Is this meant to be a new heading? What is the conclusion?

Discussion

Studies decomposing cancer mutations into a linear combination of signatures have provided invaluable insights into cancer biology (REF). Through inferring mutational signatures and latent mutational processes, researchers have gained

a better understanding of one of the fundamental driving forces of cancer initiation and development: mutagenesis.

How to leverage results from large-scale signature studies and apply them to a small set of incoming samples is a very practical problem for many researchers. Although it might seem to be a simple linear system problem at first, the core challenge is how to promote sparsity and prevent over- and underfitting. Signature studies on large-scale cancer datasets have revealed that mutational signatures are not all active in one sample or cancer type. In most tumor cases, only a few signatures prevail. A recent signature summary suggested that 2 to 13 known signatures are observed in a given cancer type [REF], which might include hundreds and even thousands of samples. Sparse solutions are biologically sound and interpretable. In addition, sparse solutions are in line with Occam's razor principle, which prefers the simplest solution that explains an observation. A desirable method should be aware of data complexity and be parameterized accordingly to avoid over- and underfitting. Finally, mutational signatures are not orthogonal due to their biological nature. Co-linearity of the signatures will lead to unstable fittings that change erratically with even a slight perturbation of the observation.

DeconstructSigs was the first tool to identify signatures even in a single tumor. This tool archives sparsity using a pre-set stopping criterion for adding new signatures in forward selection. The overly greedy nature of the stepwise feature selection is prone to eliminating valuable predictors in later steps that are correlated with previously selected ones (REF LARS). Here, we describe our newly developed sigLASSO, which provides a more mathematically rigorous alternative. Unlike deconstructSigs' approach of paving a forward selection path and fitting an unconstrained linear model at every step, sigLASSO uses the L1 norm to penalize the coefficients for signature selection, thus promoting sparsity. By fine-tuning the penalizing terms using prior biological knowledge, sigLASSO is

SEN HEAVY 3/26/2018 7:19 PM
Comment [10]: Is this the right word choice?

Shantao 3/26/2018 8:58 AM
Deleted: in

Shantao 3/31/2018 11:21 PM
Formatted: Highlight

able to further exploit previous signature studies from large cohorts and promote signatures that are believed to be active.

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

Shantao 3/26/2018 8:37 AM
Deleted: more occult

Moreover, under the current generative 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. Mutations are first divided into several signatures and then categorized further into 96 types based on the nucleotide composition. With the mutation number less than a few hundred, undersampling becomes a significant obstacle for reliable signature identification. The sampling is usually stable with abundant mutations in WGS. However, in WES, cancer samples with less than 50 mutations are common.

SigLASSO takes a conservative approach and utilizes subsampling to assess the signature inference ascertainment. In this way, the number of assigned signatures (model complexity) is informed by the complexity of the data. Likewise, sigLASSO does not specify a noise level explicitly beforehand, but instead uses cross-validation to parameterize. This is in contrast to deconstructSigs, which specifies a noise level of 0.05 to derive the cut-off of 0.06 for excluding “noise” signatures. In general, sigLASSO lets the data itself control the model complexity.

Shantao 3/26/2018 8:00 AM
Deleted: prudent

Shantao 3/31/2018 11:24 PM
Deleted: stopping

Finally, due to the colinearity 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 knowledge input could be cancer-type specific signatures or patient clinical information (e.g., smoking history or chemotherapy). We showcased the performance of sigLASSO on real cancer datasets. Although we lack the ground truth of the operative mutational signatures in tumors, we have several reasonable beliefs about the signature solution. SigLASSO produced signature solutions that are biologically interpretable, properly align with our current knowledge about mutational signatures, and well distinguish cancer types and histological subtypes.

SigLASSO exploits constraints in signature identifying and provides a robust framework to achieve biologically sound solutions. Due to the highly interdisciplinary nature of cancer signature research, 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. However, it is often arduous or impossible to quantify this value. Nonetheless, sigLASSO offers a framework that empowers researchers to use and integrate their biological knowledge and expertise into the model.

Shantao 3/31/2018 11:25 PM
Deleted: researchers

SEN HEAVY 3/26/2018 8:11 PM
Comment [11]: Interdisciplinary nature of what? Of cancer research?