

New name?

LIBRA: Likelihood-Based mutational signature Attribution in cancer genomics

Other keywords: [Jointly optimizing](#), [multinomial distribution](#), [sampling error](#), [adaptive?](#)

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 deconvolute cancer mutations by nucleotide context into a linear combination of mutational signatures. We formulated the task as a likelihood based optimization problem with L1 regularization and developed a software tool, LIBRA. First, by explicitly formulating multinomial sampling into the likelihood function and jointly optimizing the sampling process and the signature fitting, LIBRA is aware of the sampling uncertainty. It is especially pivotal in high sampling variance settings, for example, when we only observe low mutation counts in whole exome sequencing (WES). Moreover, LIBRA uses L1 regularization to parsimoniously assign signatures to mutation profiles, leading to sparse and more biologically interpretable solutions. Additionally, LIBRA integrates prior biological knowledge harmoniously by fine-tuning penalties on coefficients. Compared with hard thresholding signatures, our method leaves leeway for noise and rare signatures. Last, the model complexity is informed by the size and complexity of the data through empirical parameterizing based on performance [STL: maybe delete this sentence?]. In sum, LIBRA fits a signature deconvolution jointly with a multinomial sampling process, while using regularization to promote sparsity and interpretability. Meanwhile, this framework empowers researchers to use and integrate their biological knowledge and expertise into the model.

Introduction

- Shantao 5/12/2018 4:05 PM
Deleted: mposin
- Shantao 5/12/2018 4:05 PM
Deleted: g
- Shantao 5/12/2018 4:06 PM
Deleted: 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
- Shantao 5/6/2018 9:48 PM
Deleted: adding a
- Shantao 5/6/2018 9:48 PM
Deleted: step
- Shantao 5/6/2018 9:44 PM
Deleted: uncertainty
- Shantao 5/7/2018 12:10 AM
Deleted: mixture
- Shantao 5/6/2018 9:49 PM
Deleted: critical
- Shantao 5/6/2018 9:52 PM
Deleted: By
- Shantao 5/6/2018 9:52 PM
Deleted: assigning
- Shantao 5/12/2018 4:08 PM
Deleted: cancer genome
- Shantao 5/6/2018 9:53 PM
Deleted: the solution becomes
- Shantao 5/6/2018 9:55 PM
Deleted: sigLASSO
- Shantao 5/12/2018 4:09 PM
Deleted: into the solution
- Shantao 5/6/2018 9:54 PM
Deleted: subsetting
- Shantao 5/6/2018 9:55 PM
Deleted: before fitting
- Shantao 5/6/2018 9:54 PM
Deleted: unknown
- Shantao 5/7/2018 7:54 PM
Formatted: Highlight
- Shantao 5/6/2018 9:55 PM
Deleted: T
- Shantao 5/7/2018 7:54 PM
Deleted: using cross-validation an ... [1]
- Shantao 5/6/2018 9:55 PM
Deleted: sigLASSO
- Shantao 5/6/2018 11:57 PM
Deleted: offers a framework that

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 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 attribute 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

Shantao 5/7/2018 12:47 AM
Deleted: assign

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 a post hoc empirical pruning to achieve sparsity](#) or iterate all combinations by brute force (REF) [with a pre-fixed, small number of signatures](#). Other approaches use linear programming (REF), which is not efficient in optimization. [None of the approaches explicitly formulates the multinomial sampling process into the model](#). Here, we formulated the task as [a likelihood based/joint?](#) optimization problem with L1 regularization. First, by [jointly fitting signatures with a multinomial sampling process](#), LIBRA is aware of the sampling uncertainty. [This property](#) is especially critical in high sampling variance settings, for example, when we only observe low mutation counts in whole exome sequencing (WES). Second, [LIBRA](#) 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. [LIBRA](#) uses L1 norm, which 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, [LIBRA](#) is aware of data complexity such as mutational number and patterns in the observation. Our method is automatically parameterized [empirically on performance](#), allowing data complexity to inform model complexity. [Our](#) approach promotes result reproducibility and fair comparison of datasets.

Material and Methods

Shantao 5/7/2018 12:22 AM
Deleted: with an empirically derived stopping criterion

Shantao 5/12/2018 4:16 PM
Formatted: Highlight

Shantao 5/7/2018 12:52 AM
Deleted: explicitly adding a multinomial sampling step into likelihood function and

Shantao 5/7/2018 12:52 AM
Deleted: optimizing

Shantao 5/7/2018 12:52 AM
Deleted: sampling uncertainty and mixture

Shantao 5/7/2018 12:52 AM
Deleted: fitting

Shantao 5/7/2018 12:53 AM
Deleted: It

Shantao 5/6/2018 9:50 PM
Deleted: our sigLASSO approach

Shantao 5/7/2018 12:58 AM
Deleted: Our sigLASSO

Shantao 5/12/2018 4:18 PM
Deleted: unlike previous methods, sigLASSO

Shantao 5/7/2018 7:54 PM
Deleted: based on cross-validation

Shantao 5/12/2018 4:19 PM
Deleted: This

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 ([with](#) Frobenius norm penalty) 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. It maintains the original Frobenius norm:

$$W = \operatorname{argmin}_{W \in \mathbb{Z}^+} \|M - SW\|_2^2$$

The mutation matrix, M , contains mutations of each sample cataloged into 96 trinucleotide contexts. m_i ($i = 1 \dots n$) in M denotes the mutation count of the i^{th} category. 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.

Sampling variance

In practice, this problem is optimized on \mathbb{R}^+ [instead of integers](#) for efficiency and simplicity (REF), ignoring the discrete nature of mutation counts. [This approach essentially transforms observed mutations into a multinomial probability distribution, making model insensitive to the total mutation count.](#) Yet the total mutation count plays a critical role in inference. [Assuming](#) mutations are drawn from an underlying probability distribution (which is the mixture of several mutational signatures), the mutations follow a multinomial distribution. [The total mutation count is the sample size of the distribution, and affects the variance.](#) [For instance](#), 20 mutations of 96 categories give us very little confidence in inferring the underlying mutation distribution. If we observed 2,000 mutations, we would have [much](#) higher confidence. Methods indiscriminating these two

Shantao 5/7/2018 1:01 AM

Deleted: It

Shantao 5/7/2018 7:08 PM

Deleted: to optimization

Shantao 5/7/2018 1:03 AM

Deleted: If we assume

Shantao 5/7/2018 1:06 AM

Deleted: count

Shantao 5/7/2018 1:06 AM

Deleted: s

scenarios are clearly defective. Here, we aim to use a [likelihood-based](#) approach to acknowledge [the sampling variance and design a tool sensitive to the total mutation count](#).

LIBRA model [\[\[I still need to fix the notations...I compiled the LaTeX and pasted here as figures. Also now this a mixture of word/LaTeX\]\]](#)

We break data generation process into two parts: first, multiple mutational signatures mix together to form an underlying mutation distribution. Second, we observe a set of categorical data (mutations), which is a realization of the underlying mutation distribution. We use m_i ($i = 1 \dots n$) to denote the mutation count of the i^{th} category. \vec{p} is the underlying mutation probability distribution with p_j denote the probability of the j^{th} category.

$$L = P(\vec{p} | SW) = P(\vec{m} | \vec{p}) P(\vec{p} | SW)$$

To promote sparsity and interpretability of the solution, [LIBRA](#) uses [adds](#) an L1 norm regularizer on the weights (i.e., coefficients) of the signatures. LASSO is mathematically justified and can be computationally efficiently solved (REF). [The](#) log-likelihood looks like:

$$l \propto \sum_{i=1}^n \{m_i \log p_i - \frac{\alpha}{2} (p_i - \sum_{k=1}^K w_{ik} H_k)^2 - \lambda \sum_{k=1}^K c_k w_k\}$$

$$s.t. \forall w_k \geq 0, \forall p_i \geq 0, \sum_{i=1}^n p_i = 1$$

Here, $\alpha = 1/\sigma^2$. We [can](#) infer α from the residual errors from linear regression. λ is parameterized [empirically \(see below\)](#), \vec{c} is a vector of 30 [penalty weights](#) (c_1, c_2, \dots, c_k), 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](#). We also use \vec{c} to perform adaptive LASSO (REF) by initialize \vec{c} to $1/\beta^{\text{OLS}}$, where β^{OLS} are the coefficients from nonnegative ordinary least square. The aim is to get less biased estimator by applying [smaller penalties](#) on larger values.

Shantao 5/7/2018 1:04 AM
Deleted: likelihood based...ikelihood ... [2]

Shantao 5/7/2018 7:16 PM
Deleted: y_i in Y ... denote the ... [3]

Shantao 5/7/2018 7:15 PM
Deleted: M

Shantao 5/7/2018 7:16 PM
Formatted: Subscript

Shantao 5/7/2018 7:16 PM
Deleted: j

Shantao 5/7/2018 7:14 PM
... [4]

Shantao 5/7/2018 7:23 PM
Deleted: sigLASSO ... LIBRA uses L ... [5]

Shantao 5/12/2018 11:47 PM
Deleted: .

Shantao 5/7/2018 7:46 PM
 $l \propto \sum_{i=1}^n \{y_i \log \pi$

Deleted: Unknown

Formatted: Font:(Default) Arial

Shantao 5/7/2018 8:05 PM
 $s.t. \forall w_k \geq 0, \forall p_i \geq 0,$

Deleted: Unknown

Formatted: Font:(Default) Arial

Shantao 5/12/2018 11:48 PM
Deleted: will

Shantao 5/7/2018 7:55 PM
Deleted: by cross-validation... [6]

Shantao 5/7/2018 7:21 PM
Deleted: C... is a vector of 30 indic... [7]

Shantao 5/12/2018 11:49 PM
Moved (insertion) [1]

Shantao 5/12/2018 11:49 PM
Moved up [1]: This value should be tuned to reflect the level of confidence in prior knowledge.

Shantao 5/12/2018 11:49 PM
Deleted: can

Shantao 5/12/2018 11:49 PM
Deleted: C... to perform adaptive L... [8]

Shantao 5/12/2018 11:49 PM
Deleted: C is of the form ... β^{OLS} , w... [9]

Optimizing LIBRA

The negative log likelihood is convex in respect to both \vec{p} and \vec{w} when evaluated individually. Hence the loss function is biconvex. We optimize the function by Alternative Convex Search (ACS), which iteratively updates these two variables.

Algorithm 1 LIBRA algorithm

```

1: procedure MAINALGORITHM
2: initialization:
3:    $p_i^0 \leftarrow p_i^{MLE} = \frac{m_i}{\sum_{i=1}^n (m_i)}$ 
4:    $t \leftarrow 0$ 
5: loop:
6:    $\vec{w}^{t+1} \leftarrow \underset{\vec{w} = w_1, w_2, \dots, w_1, \dots, w_n}{\operatorname{argmax}} \sum_{i=1}^n \left\{ \frac{\alpha}{2} (p_i^{t+1} - \sum_{k=1}^K s_{ik} w_k)^2 - \lambda \sum_{k=1}^K c_k w_k \right\}$ 
                                                                                               ( $\vec{w}$ -step)
7:    $\vec{p}^{t+1} \leftarrow \underset{\vec{p} = p_1, p_2, \dots, p_i, \dots, p_n}{\operatorname{argmax}} \sum_{i=1}^n \left\{ m_i \log p_i - \frac{\alpha}{2} (p_i - \sum_{k=1}^K s_{ik} w_k)^2 \right\}$ 
                                                                                               ( $\vec{p}$ -step)

8:   if  $p^{t+1} \doteq p^t$  then
9:     break
10:   $t \leftarrow t + 1$ 
11: return  $\vec{w}^{t+1}$ 

```

To begin the iteration, we initialize \vec{m} using its maximum likelihood estimator and start with the \vec{w} -step. \vec{w} -step is a nonnegative linear LASSO regression that can be efficiently solved by glmnet (REF). λ is parameterized empirically by repeatedly splitting the nucleotide contexts into training set and testing set. We split the data set into eight subsets. Each subset contains two of every single nucleotide substitutions. We then hold off one subset as the testing dataset and only fit the signatures on the remaining ones. After circling all eight subsets and repeating the process for twenty times, we used the largest λ (which leads to a sparser solution) that gives mean square error (MSE) within one standard deviations (SD) of the minimum.

Then we use the LASSO error variance estimator to estimate λ (REF). We solve the \vec{p}^{t+1} with a Lagrange multiplier to maintain the linear summation

Shantao 5/7/2018 7:28 PM
Formatted: Font:Not Italic

Shantao 5/7/2018 7:25 PM
Deleted: M is a latent variable, thus

Shantao 5/7/2018 7:25 PM
Deleted: w

Shantao 5/7/2018 7:31 PM
Deleted: e

Shantao 5/7/2018 7:31 PM
Deleted: optimize

Shantao 5/7/2018 7:32 PM
Deleted: problem using

Shantao 5/7/2018 7:25 PM
Deleted: expectation maximization (EM).

Shantao 5/13/2018 12:11 AM
Deleted: -step: ... [10]

Shantao 5/7/2018 8:03 PM
Deleted: $m_i = y_i / \sum y_i$

Shantao 5/7/2018 7:53 PM
Deleted: M

Shantao 5/7/2018 7:53 PM
Deleted: M

Shantao 5/7/2018 7:55 PM
Deleted: by 12-fold (since 12 is a divisor of 96) cross-validation

Shantao 5/13/2018 12:13 AM
Deleted: At every step, w

Shantao 5/7/2018 8:00 PM
Deleted: hold off eight trinucleotide context

Shantao 5/7/2018 8:00 PM
Deleted: s

Shantao 5/7/2018 7:58 PM
Deleted: 88 trinucleotide contexts

Shantao 5/7/2018 8:01 PM
Deleted: We

Shantao 5/7/2018 7:59 PM
Deleted: three

Shantao 5/7/2018 8:02 PM
Deleted: M^{t+1} using

constrain $\sum_{i=1}^n p_i = 1$. The nonnegative constrain of p_i is satisfied in only retain a nonnegative root of the solution (see Appendix?).

The key step is the \vec{p} -step. In this step, we try to estimate \vec{p} that optimizes the multinomial likelihood *while* constrain it not too far away from the fitted $\hat{\vec{p}}$. If we only use the point MLE of \vec{p} and do not perform the \vec{p} -step, the model assumes the sampling is perfect and becomes insensitive to the total mutation counts. The trade-off in \vec{p} -step between the multinomial likelihood and the L2 loss reflects the sampling error. The sampling size (sum of m_i), the goodness of signature fit (as reflected in α) and the overall shapes of $\hat{\vec{p}}$ all affect the tension between sampling and linear fitting.

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 addictive Gaussian noise was simulated at various levels with a positive normal distribution on 25% trinucleotide contexts. Then, we summed all the signatures and noise to form a mutation distribution. We sample mutations from this distribution with different mutation counts.

We ran deconstructSigs according to the original publication (REF) and LIBRA 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 (support recovery).

Illustrating on real datasets

Shantao 5/13/2018 12:15 AM

Deleted: This process is iterated until convergence.

Shantao 5/13/2018 12:26 AM

Deleted: uniform

Shantao 5/13/2018 12:26 AM

Deleted: 96

Shantao 5/13/2018 12:27 AM

Deleted: randomly drew

Shantao 5/13/2018 12:27 AM

Deleted: sigLASSO

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.

LIBRA software suite

[LIBRA](#) accepts processed mutational spectrums. We provided simple scripts to help parse mutational spectrums from VCF files. [LIBRA](#) allows users to specify biological priors (i.e., signatures that should be active or inactive), subsampling steps, and the subsampling cutoff. [LIBRA](#) uses 30 COSMIC signatures by default. Users are also given the option to supply customized signature files.

[LIBRA](#) is computationally efficient; using default settings, the program can successfully decompose a whole genome sequenced (WGS) cancer sample in [a few seconds](#) on a regular laptop (3 GHz i7 CPU, 16 GB DDR3 memory).

We have released [LIBRA](#) as an R package. The updated code is also available on GitHub (<https://github.com/Shantaol/LIBRA>).

Results

[LIBRA is aware of the sampling variance](#)

[Jointly optimizing both sampling process and signature fitting, LIBRA is aware of the sampling variance and infers an underlying mutational context distribution p. As shown in Figure 2A, the underlying latent distribution is optimized in respect to both sampling likelihood and the linear fitting of signatures. In low mutation](#)

Shantao 5/13/2018 12:28 AM

Deleted: SigLASSO

Shantao 5/13/2018 12:28 AM

Deleted: SigLASSO

Shantao 5/13/2018 12:28 AM

Deleted: SigLASSO

Shantao 5/13/2018 12:28 AM

Deleted: SigLASSO

Shantao 5/13/2018 12:28 AM

Deleted: LASSO

Shantao 5/13/2018 1:59 AM

Deleted: less than a minute

Shantao 5/13/2018 12:28 AM

Deleted: SigLASSO

Shantao 5/13/2018 12:28 AM

Deleted: SigLASS

Shantao 5/13/2018 12:28 AM

Deleted: 0

Shantao 5/13/2018 2:08 AM

Formatted: Font:Not Bold

counts, the uncertainty in sampling increases and thus the estimated underlying distribution goes closer to the least square of signatures. In contrast, when the total mutation counts is high, the estimate of the distribution is closer to the MLE.

We subsampled a real WGS cancer (papillary renal cell carcinoma, TCGA-B9-A44B, Figure 2B, REF) with various sample sizes. When the sample size is small (<100), high uncertainty in sampling pushed the inferred underlying mutational distribution p far from the MLE in trade for better signature fitting. When the sample size increases, lower variance in sampling dragged p close to MLE and forced the signatures to fit with larger error.

Meanwhile, deconstructSigs is not sensitive to mutation counts. It exhibits higher fitted coefficient variance in low mutation counts subsampling.

Performance on a simulated dataset

We first evaluated LIBRA on a simulated dataset. Both LIBRA and deconstructSigs performed better with higher mutation number and lower noise (Fig. 3). Overall, LIBRA achieved lower mean square error (MSE) than deconstructSigs.

A decrease in mutation number leads to an increase of uncertainty in sampling, which is mostly negligible in the high mutation scenarios. As expected, the MSE jumped to the 0.05-0.3 range for noiseless, low and high noise setups when the mutation number was low. Thus, the error is dominated by undersampling rather than embedded noise. LIBRA consistently performed better than deconstructSigs in low mutation counts.

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

- Shantao 5/13/2018 2:02 AM
Deleted: SigLASSO
- Shantao 5/13/2018 2:02 AM
Deleted: sigLASSO
- Shantao 5/13/2018 2:02 AM
Deleted: 2
- Shantao 5/13/2018 2:02 AM
Deleted: For a program that recovers all signatures perfectly but is oblivious to 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.
- Shantao 5/13/2018 2:04 AM
Deleted: .
- Shantao 5/13/2018 2:05 AM
Deleted: 1
- Shantao 5/13/2018 2:06 AM
Deleted: both

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.

WGS scenario using renal cancer datasets

We benchmarked the two methods using 35 WGS papillary kidney cancer samples (Fig. 4, REF). The median mutation count was 4,528 (range: 912-9,257). We found that without prior knowledge, both [LIBRA](#) and [deconstructSigs](#) showed high contributions from signature 3 and [5. deconstructSigs also assign a high proportion signature 8, 9 and 16 .](#) Signatures 3, 8, 9 and 16 were not found to be active 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.

When [LIBRA](#) took into account the prior knowledge of active signatures, [the proportion of backbone signature 5 increased to about 75%](#), 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

Shantao 5/13/2018 4:52 AM
Deleted: 3

Shantao 5/13/2018 4:34 AM
Deleted: sigLASSO

Shantao 5/13/2018 4:35 AM
Deleted: 8

Shantao 5/13/2018 4:35 AM
Deleted: ;

Shantao 5/13/2018 4:36 AM
Deleted: these s

Shantao 5/13/2018 4:36 AM
Deleted: in

Shantao 5/13/2018 4:51 AM
Deleted: sigLASSO

Shantao 5/13/2018 4:51 AM
Deleted: the assignment increased to around 70% in most cases.

Shantao 5/13/2018 4:51 AM
Deleted: T

Shantao 5/13/2018 4:52 AM
Deleted: was

Shantao 5/13/2018 4:52 AM
Deleted: signature 5

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.

...

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.

2.3 Performance on 8,892 TCGA samples

....

LIBRA is computationally efficient

LIBRA iteratively solves two convex problems. The w-step can be solved using a very efficient coordinate descent algorithm (glmnet). The p-step is solving a set of quadric equations. We observed empirically the solution quickly converged in a few iterations even with extremely low mutation numbers (~10). In constast, deconstructSigs uses simple binary search to fit coefficients. Every step it loops through all signatures and perform binary search on each one of them. deconstructSigs scales poorly when the number of signatures goes up.

We profiled ...

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

Shantao 5/13/2018 3:07 AM

Formatted: Font:Bold

Shantao 5/13/2018 4:03 AM

Formatted: Highlight

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 prevent over- and underfitting on only one single sample, often, with very few mutations (especially in WES) and promote sparsity. First, 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, sampling variance becomes a significant factor in reliable signature identification. Therefore, the fitting scheme should be aware of the sampling variance, which is especially pronounced in low mutation count scenarios (WES or cancer types with low mutation burden). A designed tool should be able to attribute the signatures by flexibly inferring the underlying true mutation distribution given the sampling variance and the signature fitting performance. Second, the solution should be sparse. 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. Third, a desirable method should be aware of data complexity and be parameterized accordingly to achieve the optimum fitting. 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.

Shantao 5/13/2018 1:09 AM

Deleted: promote sparsity and

Shantao 5/13/2018 1:27 AM

Moved (insertion) [2]

Shantao 5/13/2018 1:28 AM

Deleted: undersampling

Shantao 5/13/2018 1:28 AM

Deleted: obstacle for

Shantao 5/13/2018 1:27 AM

Deleted: The sampling is usually stable with abundant mutations in WGS. However, in WES, cancer samples with less than 50 mutations are common.

Shantao 5/13/2018 1:27 AM

Deleted: -

Shantao 5/13/2018 1:10 AM

Deleted: A

Shantao 5/13/2018 1:10 AM

Deleted: avoid over- and underfitting

DeconstructSigs was the first tool to identify signatures even in a single tumor.

This tool uses forward selection and archives sparsity by a post hoc pruning with a preset 6% cut-off. First, DeconstructSigs is insensitive to the total mutation counts. The mutation spectrum is normalized before fitting thus makes mutation counts irrelevant. Moreover, 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 LIBRA, which jointly optimized the sampling process and an L1 regularized signature fitting. By explicitly formulating a multinomial sampling likelihood into the optimization, LIBRA is aware of the sampling variance. Meanwhile, unlike deconstructSigs, which paves a forward selection path and fits an unconstrained linear model at every step, sigLASSO uses the L1 norm to penalize the coefficients, thus promoting sparsity. By fine-tuning the penalizing terms using prior biological knowledge, sigLASSO is able to further exploit previous signature studies from large cohorts and promote signatures that are believed to be active.

By jointly optimizing a “mutation sampling” process enables LIBRA to be aware of the sampling variance. We demonstrated by additionally modeling a multinomial sampling process, LIBRA is able to achieve better signature attribution, especially in low mutation counts cases. In cancer research, WES data is abundant but it also suffers severely from undersampling in signatures attribution. In these cases, LIBRA can generate more reliable and robust solutions. Moreover, we believe formulating mutation by a multinomial process can have further implications in background mutation rate modeling (REF encodec?).

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

Shantao 5/13/2018 1:11 AM

Deleted: using a pre-set stopping criterion for adding new signatures in forward selection

Shantao 5/13/2018 1:11 AM

Formatted: Font:Italic

Shantao 5/13/2018 1:14 AM

Deleted: T

Shantao 5/13/2018 1:15 AM

Deleted: our newly developed sigLASSO

Shantao 5/13/2018 1:16 AM

Deleted: provides a more mathematically rigorous alternative

Shantao 5/13/2018 1:24 AM

Deleted: U

Shantao 5/13/2018 1:18 AM

Deleted: ' approach of

Shantao 5/13/2018 1:24 AM

Deleted: ing

Shantao 5/13/2018 1:24 AM

Deleted: fitting

Shantao 5/13/2018 1:24 AM

Deleted: for signature selection

Shantao 5/13/2018 1:53 AM

Formatted: Highlight

Shantao 5/13/2018 1:53 AM

Formatted: Highlight

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).

Moreover, [LIBRA](#) does not specify a noise level explicitly beforehand, but instead [empirically tunes parameters based on model performance](#). 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, [LIBRA](#) lets the data itself control the model complexity.

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, LIBRA 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 LIBRA 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. LIBRA produced signature solutions that are biologically interpretable, properly align with our current knowledge about mutational signatures, and well distinguish cancer types and histological subtypes.

[Due to the highly interdisciplinary nature of cancer signature research, identifying signatures in cancer samples is a challenging task. In this work, we introduced LIBRA, which exploits constraints in signature identifying and provides a robust framework to achieve biologically sound solutions. It jointly optimized a sampling](#)

Shantao 5/13/2018 1:27 AM

Moved up [2]: 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.

Shantao 5/13/2018 1:27 AM

Deleted: [...](#) [1]

Shantao 5/13/2018 1:30 AM

Deleted: uses cross-validation

Shantao 5/13/2018 1:30 AM

Deleted: to

Shantao 5/13/2018 1:30 AM

Deleted: parameterize

Shantao 5/13/2018 1:36 AM

Deleted: .

Shantao 5/13/2018 1:54 AM

Moved (insertion) [3]

process with an L1 regularized signature fitting. Additionally, LIBRA is also able to empower researchers to use and integrate their biological knowledge and expertise into the model.

Shantao 5/13/2018 1:54 AM

Moved up [3]: Due to the highly interdisciplinary nature of cancer signature research, identifying signatures in cancer samples is a challenging task.

Shantao 5/13/2018 1:54 AM

Deleted: 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, LIBRA offers a framework that