Discovering novel mutational signatures by compound latent Dirichlet allocation with variational Bayes inference

Prepared by: Mina Shaigan

Abstract

Understanding DNA damage and repair processes operating during the cellular lineage between the fertilized eggs and the cancer cells is a major problem in cancer biology. Somatic mutations caused by these processes are associated with DNA replication or exposure to environmental and lifestyle factors. Knowing the ongoing mutational processes in human cancer can be used in personalized and targeted therapies, prognosis, and oncological research.

Computational methods have identified 30 mutational signatures of mutational processes active in human cancers that biologist experts have validated. Each signature is a pattern of mutation types. However, different signatures are obtained from other tissues that may have the same primary lesion. To determine a mutational signature appropriate to a mutational process, it is necessary to merge the signatures obtained from all the primary lesions in which the mutational process has operated. Samples of each primary lesion should not be considered separately but together. Previous models have examined the types of primary lesions separately, which is inefficient for the comparative analysis of samples.

I propose the Compound Latent Dirichlet Allocation model in which the sample set can be divided into primary lesion subgroups. Some mutational signatures can be common between subgroups, but the proportion of signatures for each type of primary lesion is different. This model can capture the characteristics of each primary lesion. With analysis of 84,729,690 somatic mutations from 4,645 whole cancer genomes, published by the COSMIC database, I characterized single base substitution and insertion and deletion of mutational signatures and their exposures in each genome. Also, the model allows us to examine the relative contributions of each of the mutational signatures in each primary lesion. I utilize a variational Bayesian inference to identify latent variables that enables us to select a model with the most probable number of mutational patterns. By applying our model for real mutation data, I decipher new mutational signatures previously unknown.

Keywords:

Mutational process, Cancer, Mutation signature, Bayesian inference

TABLE OF CONTENTS

1. Introduction

1.1. The context of problem

The genome of a cancer cell consists of somatic mutations that result from DNA damage and repair processes from birth to the emergence of cancer cells, leading to specific mutation patterns. I know that each mutational process leads to certain mutations, and when these mutations are more preferred to a process, those mutations are called mutational signatures. The accumulation of mutations from birth to the present day is thought to be the result of a combination of mutational signatures. Identifying mutation signatures from cancer-derived somatic mutational signatures is an important task in describing carcinogenic mechanisms and may also be used as biomarkers for early detection of cancer.

1.2. The problem definition

Each patient's cancer cell could be exposed to different intensities and duration of mutational processes. The whole-genome sequence as a mutation catalog is a combination of hidden patterns of mutational processes. These latent patterns are called mutational signatures. Each mutational signature as a distribution of mutations corresponds to one mutational process.

Input: mutation catalogs, a set of mutations (e.g. a substitution from A to C with [A> C]) of cancer genome sequences.

Output: mutational signatures, distribution of mutations for each mutational process, and intensities of exposure of mutational processes for each cancer genome.

1.3. The background of problem

Several different procedures, such as NMF-based and statistical model-based, have been used to decipher these patterns separately. In the non-negative matrix factorization (NMF) method the mutation catalogs matrix decompose into two matrices, mutational signatures matrix, and exposures matrix so that the Frobenius norm between them becomes minimized [1]. Since the number of mutation signatures is unknown, different approaches like clustering and empirical Bayesian are used for model selection [4]. In contrast, statistical approaches using a probabilistic framework like EM [6], mixed-membership models [5], and LDA [7] can perform better in the face of biases [2].

1.4. Purposes

Although these methods have a high ability to find mutational signatures, they have some limitations. In the previous methods, mutation catalogs for each cancer type are examined separately. In this way, one might have to align several mutational signatures to one mutational process so that they can be different. Moreover, some information will be lost, and for a small dataset, this will cause a problem. To address these problems, Matsutani and his colleagues use parallelized latent Dirichlet allocation model, SigProfiler software clusters mutational signatures of all cancer types based on “reproducibility” as a distance measure, and SignatureAnalyzer applied Bayesian non-negative matrix factorization.

I propose the compound latent Dirichlet allocation model (cLDA) that considers cancer type as prior knowledge on mutation catalogs in the modeling framework. cLDA takes a latent mixture of mutational signatures for genome samples in all cancer types and uses it to extract mutational signatures for each genome sample. cLDA can also obtain the distribution of mutational signatures in each cancer type. Hidden variables of this model are derived from the posterior probability. I approximate intractable posterior probability by using a variational bayesian method.

This method can be evaluated on simulated data according to cLDA model by changing the total number of samples, the number of mutations in each sample and the model parameters with the COSMIC known signatures. I also implement our method for predicting actual signatures over mutation catalogs in the COSMIC database to search for new signatures. To evaluate and compare our method, I use the cosine distance and the similarity of mutation patterns (by comparing the number of mutations for each mutation type).

2. Research methods and implementation steps

2.1. Collection and preparation of problem data

First, DNA of patients is extracted from cancerous tissue and healthy tissue (skin tissue or healthy part of cancerous organ). I only consider people who have more than 70% of their cancer cells. The number of known cancer genomes in the world is growing due to recent advances in next generation sequences (NGS). The mean sequence coverage is 40.4 fold for cancer samples and 30.2 fold for healthy samples. CaVEMan2 method can be used to call somatic mutations (detect the presence of mutations in the results of NGS experiments). With a specific cancer genome, I obtain a set of mutations in the genome. These mutations are available from targeted and extensive parts of the genome in the COSMIC database. This database includes sample name, primary site, primary histology, mutation CDS, mutation description, mutation genome position, mutation strand, and etc. The number of these mutations is 84,729,690 and has been extracted from 2780 whole cancer genomes. There are 4 types of mutations: base substitutions, small insertions and deletions (indels), rearrangements, and copy number alterations. A base substitution from C to A is displayed as [C> A]. Each nucleotide can be replaced with 3 other types of nucleotides. If a nucleotide is substituted in a DNA sequence, a substitution also takes place in proportion to that type of substitution on the complementary strand. As a result, I have 6 types of base substitution: [C > A], [C > G], [C > T], [T > A], [T > C], [T > G].

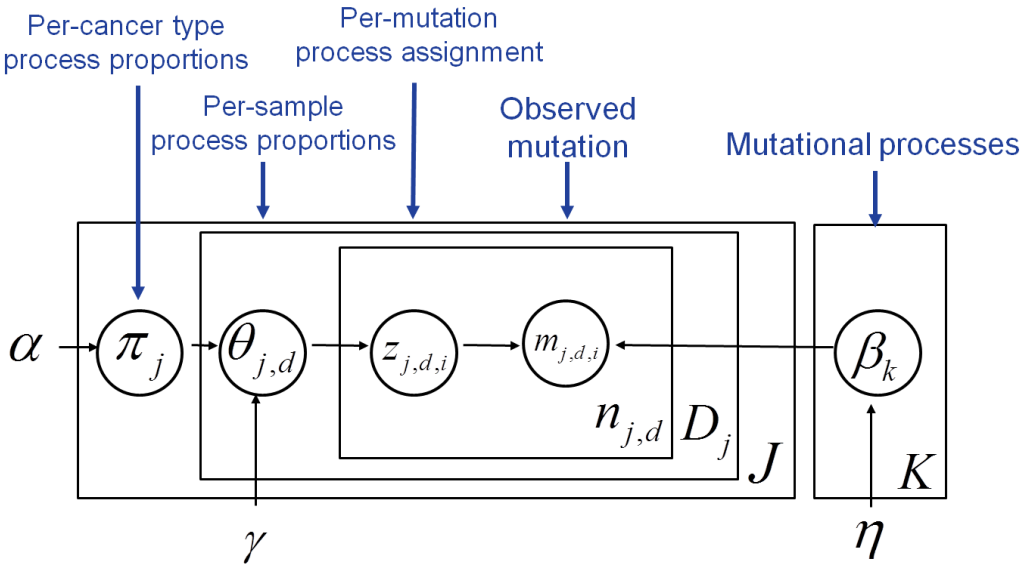
2.2. Method development and implementation

Mutations are modeled using a probabilistic generative model called compound latent Dirichlet allocation. This model has already been proposed for the topic modeling with different categories [3]. In this case, for the documents of each category, a mixture of hidden common topics at the category level is assumed. This mixture is used in each category as a scale for extracting a mixture of hidden topics for each document. Each word is generated from a document of a category by a topic that has a specific distribution of words. In this study, the categories, vocabularies, words, and topics corresponding to the cancer types, mutation types, mutations, and mutational signatures, respectively.

Iuse the following symbols:

- $m_{jdi} \in \{A[C > A]A, A[C > A]C, \ldots, T[T > G]T\}$, the ith mutation in dth genome sample in jth cancer type
- J, number of cancer types; $j(1 \geq j \geq J)$is an index for a cancer type;
- D_j, number of samples of jth cancer type;
- n_{jd}, number of mutations of dth sample of jth cancer type; $i(1 \geq i \geq n_{jd})$is an index for a mutation;
- V, number of mutation types; $v(1 \geq v \geq V)$is an index for a mutation type;
- K, number of mutational signatures; $k(1 \geq k \geq K)$is an index for a mutational signature.
- $\theta_{jd} = \{\theta_{jdk}\}_{k=1}^{K}$, is the parameter of the multinomial distribution of mutational signatures for each sample d of jth cancer type; Where θ_{jdk} represents the activity of the kth mutational process in the dth genome sample of the jth cancer type.
- $\beta_k = \{\beta_{kv}\}_{v=1}^{V}$, is the parameter of the multinomial distribution of mutations for each mutational signature k; Where β_{kv} represents the proportion of vth mutation type in kth mutational signature.
- $\pi_j = \{\pi_{jk}\}_{k=1}^{K}$, is the parameter of the Dirichlet distribution of mutational signatures in a jth cancer type.

In this model, the number of mutational signatures K is assumed to be known.
The graphical model representation of the compound latent Dirichlet allocation for this generative process is given in figure below.



In this study, variational Bayesian inference is used because the evaluation function (called the variational lower bound) is used for model selection (estimating the number of hidden variables).

2.3. Method evaluation

I use cosine similarity to compare our predicted mutational signatures with known ones. This similarity is obtained by total multiplication distributions of mutation types in predicted and known mutational signatures by multiply sum distributions of mutation types in predicted and known mutational signatures

$$\text{similarity} = \frac{\phi_k \cdot \varphi_l}{\|\phi_k\| \|\varphi_l\|} = \frac{\sum_{v=1}^{V} \phi_{kv} \varphi_{lv}}{\sqrt{\sum_{v=1}^{V} \phi_{kv}^2} \sqrt{\sum_{v=1}^{V} \varphi_{lv}^2}}$$

One of the common uses of cosine distance is to match similar texts based on counting the most common words between two texts. Because in this study the size of the mutation types is the same for every signature, as the probability of common mutations increases, it can capture characteristic peaks in a mutation distribution.

3. Results and discussion

To ensure the performance of the compound latent Dirichlet allocation model, I performed experiments with a simulated data set. In this analysis, the mutation data are simulated based on the generative process of the cLDA model, in which the known mutational signatures 1,2,…, 10 of the COSMIC database are taken.

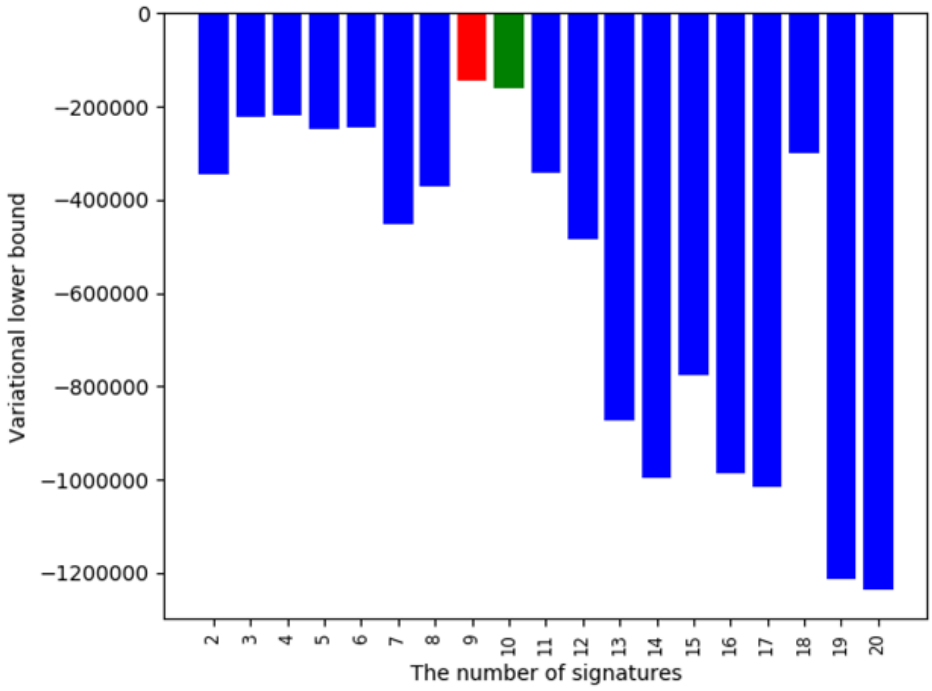
By changing the number of cancer types, the number of genomes for one cancer type, the number of mutations per genome, and the hyperparameters, I generated various simulated data to determine under what conditions the model can accurately predict the original signatures.

For K values from 2 to 20, the model was implemented and the optimal number of mutational signatures was estimated by Bayesian inference.

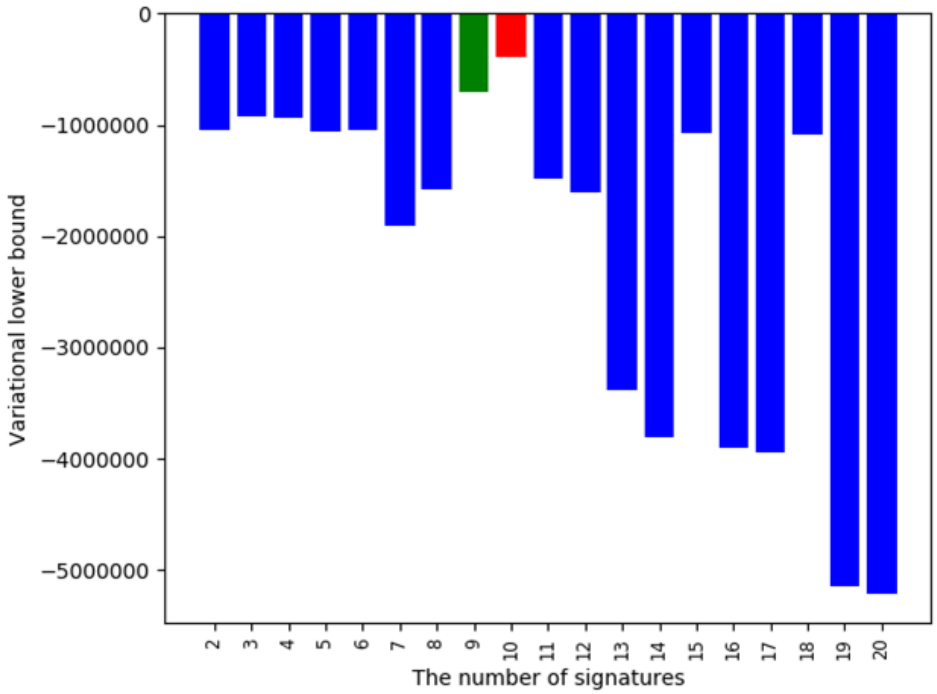
The table below shows the simulation results.

J	α	γ	# predicted signature		
			D=20	D=30	D=100
1	0.1	0.1	11	13	9
		1	6	8	9
4	1	0.1	11	13	10
		1	7	8	10

In the figure below, the lower bound values change for K = 2,…, 20 and J = 1.



In the figure below, the lower bound values change for K = 2,…, 20 and J = 4, and the mutational signatures for K = 10 , n_{jd}= 2000, and D_j= 100 has been estimated. These results clearly mean that the variational lower bound to the cLDA model has successfully estimated the correct number of mutational signatures and mutation distribution for most conditions.



To search for new signatures, I applied our method over mutation catalogs of the COSMIC database to predict actual mutational signatures. Mutational processes vary depending on the cancer type, so I separated mutation catalogs for each type of cancer and then used them to learn the cLDA model.

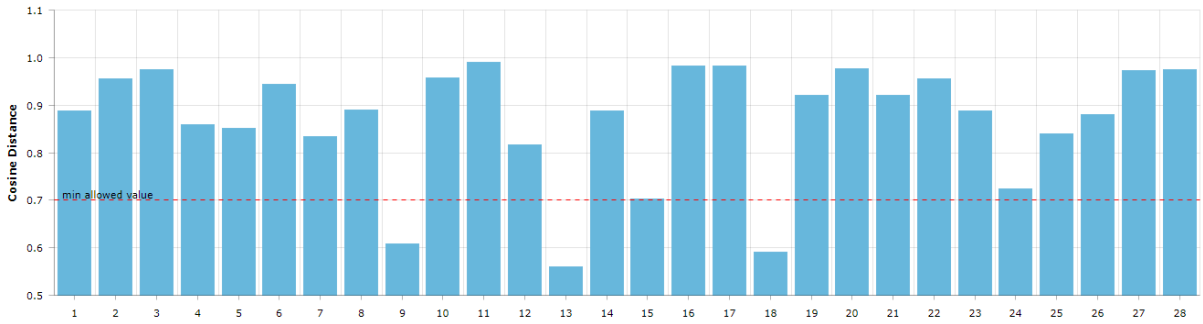
According to the results of the simulated data, if the number of genomes of a cancer type is less than 25, or if the number of mutations of a genome is less than 400, Ido not consider that type of cancer in the model and remove that genome from the model. As a result, 1607 genomes have been studied in 12 categories of cancers (breast, uterus, colon, liver, lung, esophagus, prostate, skin, soft tissue such as lymph vessels, stomach, upper gastrointestinal tract, and urinary tract).

Finally, I applied our Dirichlet allocation method to the mutation dataset and determined the mutational signatures using Bayesian inference.

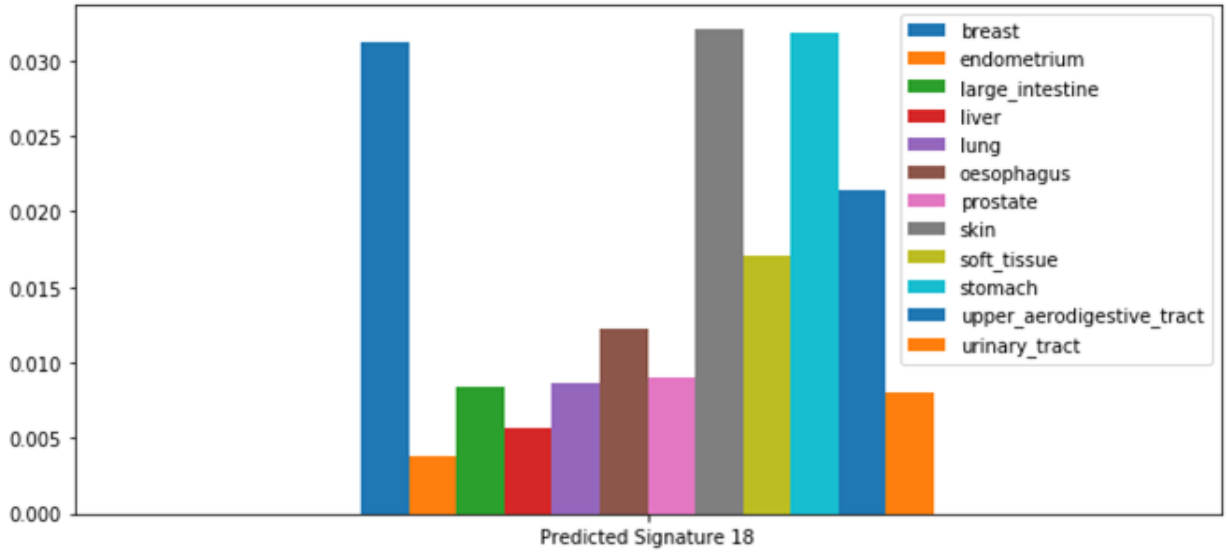
By matching the predicted signatures with the known COSMIC signatures, I obtained a mutational signature that could be new, or otherwise expand our understanding of the known COSMIC signatures.

If the cosine similarity of each predicted signature to a known COSMIC signature is greater than 0.7, I consider the two signatures to be the same. The following figure shows

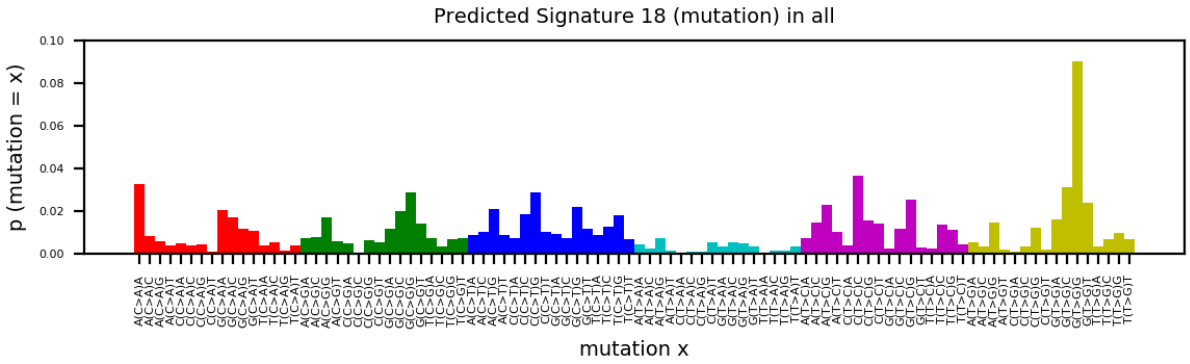
the cosine similarity of each predicted signature to the most similar known COSMIC signature.



The predicted mutation 18 signature is more common in breast, skin, and stomach cancers.



The characteristic patterns of this mutation signature are X [T> C] G (X = A, T, G, C) as well as G [T> G] G, which may be influenced by an unknown mutational process. The most similar mutational signature known to this signature is the mutational signature 3 of COSMIC, and the cosine similarity between them is relatively small (0.591413). The known mutational signature 3 does not have the pattern G [T> G] G.



By examining the signatures proportion obtained for each cancer type and comparing it with the tissue distribution provided for each known COSMIC signature, it can be seen that the cLDA model has been able to well observe the relationship between different cancer types. As a result, by analyzing the mutational signatures proportion between different types of cancer, the function of mutational processes can be understood better.

4. Conclusion

What is presented in this study, as described, is an attempt to use a model to decipher mutation patterns in the whole cancer genome. I estimated the mutational signatures using the compound latent Dirichlet allocation model, which is a generative model.

The advantage of this model is the mathematical criterion by which the model with the optimal number of mutational signatures can be selected. It also made it possible to analyze all cancer genomes together by classifying mutation data by type of cancer; Without increasing the complexity of the data set. As a result, for each mutational process, a corresponding mutational signature is found.

Also, by using this model, the distribution of mutational signatures in each type of cancer was determined. As a result, Ican examine the impact of each mutational signature on each type of cancer and gain a better understanding of the function of mutational processes.

By comparing the predicted mutational signatures with the known COSMIC mutational signatures, a new one was found that was continuously active in several genomes.

Cancers are caused by somatic mutations. Thus, mutational signatures analysis provides an important insight into cancer development through a comprehensive description of mutational processes. For conceptual and practical simplicity, it was assumed that there was a single signature associated with each mutational process. It is possible to find different mutational

signatures for a process in different tissues. Due to the differences in tissues, Iadd other features of the mutation types, such as the distance of succession mutations, transcriptional strand bias, the level of gene expression, the distance from transcriptional strand bias start site, and the propagation of mutations in genes encoding proteins.
The comprehensive collection of mutational signatures found in human cancer is the foundation for future research on (i) geographical and temporal differences in cancer incidence to elucidate underlying differences in aetiology, (ii) the mutational processes and signatures present in normal tissues and caused by non-neoplastic disease states, (iii) clinical and public health applications of signatures as indicators of sensitivity to therapeutics and past exposure to mutagens, and (iv) mechanistic understanding of the mutational processes underlying carcinogenesis [8].

5. References

[1] Alexandrov, L. B., Nik-Zainal, S. and Wedge, D. C. et al. (2013) Signatures of mutational processes in human cancer. Nature, 500, 415-421.

[2] Matsutani, T., Ueno, Y., Fukunaga, T. and Hamada, M. (2019) Discovering novel mutation signatures by latent Dirichlet allocation with variational Bayes inference. Bioinformatics, 35, 4543-4552.

[3] George, C. P., Xia, W. and Michailidis, G. (2019) Analyses of multi-collection corpora via compound topic modeling. in Machine Learning, Optimization, and Data Science (Nicosia, G., Pardalos, P., Umeton, R., Giuffrida, G. and Sciacca, V. eds.), (Cham), 205–218, Springer International Publishing.

[4] Rosales,R.A. et al. (2017) SigneR: an empirical Bayesian approach to mutational signature discovery. Bioinformatics, 33, 8–16.

[5] Shiraishi,Y. et al. (2015) A simple model-based approach to inferring and visualizing cancer mutation signatures. PLoS Genet., 11, e1005657.

[6] Fischer,A. et al. (2013) Emu: probabilistic inference of mutational processes and their localization in the cancer genome. Genome Biol., 14, R39.

[7] Blei,D.M. et al. (2003) Latent dirichlet allocation. J. Mach. Learn. Res., 3, 993–1022.

[8] Alexandrov,L. et al. (2018) The repertoire of mutational signatures in human cancer. BioRxiv, 322859.