Extraction of Latent Probabilistic Mutational Signature in Cancer Genomes

Yuichi Shiraishi[1,*], Georg Tremmel[1], Satoru Miyano[1], Matthew Stephens[2,3]

1 Laboratory of DNA Information Analysis, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan

2 Dept. of Human Genetics, University of Chicago, Chicago, Illinois, United States of America

3 Dept. of Statistics, University of Chicago, Chicago, Illinois, United States of America

∗ E-mail: yshira@hgc.jp

Abstract

Thanks to the advances in recent high throughput sequencing technologies, a massive amount of somatic mutations from cancer genome sequencing data become available. Accordingly, it becomes possible to detect characteristic patterns of somatic mutations or "mutation signatures" at an unprecedented resolution with the expectations of revealing novel causes and mechanisms of tumorigenesis.

Several statistical approaches for extracting characteristic mutation signatures have been proposed. However, in previous approaches, since the number of parameters increases exponentially as more contextual factors are taken into account, we could not treat many contextual factors all together because of instability of estimation results. Furthermore, interpretation of mutations signatures of huge dimensional vectors is often troublesome.

In this paper, we propose a novel approach based on hierarchical probabilistic modeling. The proposed approach reduces the number of parameters by the independence assumption on each factor so that we can obtain more robust and interpretable estimates. Using synthetic and real data, we demonstrate that the proposed approach can not only give highly robust estimates, but also capture novel characteristics such as base frequency at the two base 5' to the mutated sites. In addition, we clarify the relationships between the proposed approach and the "mixed-membership models", that have been actively studied in statistical machine learning and statistical genetics community. Recognizing these relationships will help us to develop a grasp of mutation signature extraction problems, and will allow us to utilize a lot of techniques accumulated on other fields to further improve the statistical methods.

Finally, we have prepared an R package of the proposed approach (probabilistic mutation signature, pmsignature), which is available at `https://github.com/friend1ws/pmsignature`.

Author Summary

It is known that the pattern of somatic mutations depends on cancer types and even individuals within the same caner type. For example, C > A mutations are frequent in lung caner whereas C > T and CC > TT mutations are frequent in skin cancer, and their patterns are usually associated to some kinds of carcinogens. However, since the cost and throughput of sequencing technologies were limited, high-resolution analysis of mutation patterns was not possible.

With the coming of high-throughput sequencing technologies, we can now obtain a massive amount of somatic mutation data from cancer genomes, and there is a great possibility of detecting novel mutation patterns leading to identification of novel carcinogens and obtaining better classification of cancer genomes based on mutation patterns differences. On the other hand, there is an increasing need for novel statistical method for mutation patterns analysis from vast amount of somatic mutation data.

In this paper, we provide a novel statistical tools for clarifying characteristic mutation patterns from massive amounts of cancer somatic mutation data, which will give us more robust and interpretable

mutation patterns compared to previous approaches. We demonstrate the efficiency of our method using simulated and real data.

Introduction

Cancer is a genomic disease. As we lead a life, DNAs within the cells throughout our body acquire a number of random somatic mutations mainly caused by DNA replication errors and exposures to mutagens such as chemical substances, radioactivities and inflammatory reactions. Although most mutations are harmless (called "passenger mutations"), a small portions of mutations at some specific sites in cancer genes (called "driver mutations") confer growth activities to the cells having them over other cells, allowing such as autonomous proliferation and tissue invasion, and contribute to oncogenesis [1]. The main goal in cancer genome study has been to find the driver mutations to understand the mechanism of cancer development. On the other hand, even passenger mutations can also give us an important information, because they often show specific mutation patterns, or "mutation signatures", which reflect driving forces causing somatic mutations. Classical analyses of mutation patterns have revealed a number of relationships between mutation patterns and carcinogens. For example, C > A mutations at CpG sites are abundant in lung cancers with smoking history, and these are caused by benzo(a)pyrene included in tobacco smoke [2]. Also, C > T and CC > TT mutations are abundant in ultraviolet-light-associated skin cancers, and these are caused by pyrimidine dimers as a result of ultraviolet radiation [3].

There were several problems in these classical studies [2, 3]. Due to limited sequencing throughput, they mostly aggregated mutations collected from each individuals over the same cancer types focusing on a few cancer genes such as TP53 where high mutation frequencies could be expected, and compared mutation patten profiles among different cancer types. However, many of those mutations in those cancer genes are considered to be driver mutations adding selective activities to the cells, which can give biased mutation profiles from those purely reflect driving forces of mutations. Furthermore, the mutation pattern profile not for each cancer type but for each individual could not be explored in the above approach.

Nowadays, recent advancement in high-throughput sequencing technologies provides vast amounts of passenger mutations and great chances for not only identification of novel driver mutations but also investigation of sample-by-sample mutation patterns in an unbiased way. A study using 21 breast cancer samples identified the association between C > [AGT] mutations at TpC sites, which is later proved to be caused by APOBEC proteins, and the novel phenomenon called *kataegis* [4]. Moreover, the recent landmark study revealed a landscape of mutation signatures using from 7,034 primary cancer samples of 30 different classes [5]. Furthermore, it is expected that detection of novel mutation signatures and associated mutagens can lead to identification of novel mutagens and prevention of cancer.

At the same time, for extracting prominent mutation signatures from vast amounts of somatic mutation data, there is an increasing need for novel latent variable modeling approaches. Currently, only a few approaches have been proposed. In [6], nonnegative matrix factorization (NMF, [7]) is used for the matrix whose elements represent frequencies of mutation patterns for each sample. In [8], the number of each mutation patterns is assumed to be generated by Poisson distribution whose parameters are linear combinations of latent mutational signatures.

One of the major problems of the previous approach is that the number of parameters in the model grows exponentially with the number of contextual factors to take into account. Many of recent approaches take into account just immediate 5' and 3' bases as contextual factors in addition to substitution patterns. However, when considering two 5' and 3' bases to the mutated sites, the estimated mutation signatures often become unstable due to the curse of high dimensionality, despite there is a potential need to observe two base 5' and 3' positions to the mutated sites [9]. Also, it is often very hard to interpret and extract important characteristics from the estimate with high-dimensional parameters. Furthermore, considering that latent variable modeling have enormous number of researches in statistics and machine learning, there is a great possibility that we can further elaborate statistical methods for mutational signature

extraction problems.

In this paper, we present a novel probabilistic approach. To avoid the problem of the exponential increase of the number of parameters, we assumed independence among contextual factors, which can give more interpretable and robust estimation. We demonstrate that independence assumption can robustly capture known mutational signatures with additional contextual informations.

In addition, even though it has not been widely discussed, the problem of identifying mutation signatures is closely related to "mixed-membership models" in other fields, such as "population admixture" models [10] in the analysis of population structure, and the "latent dirichlet allocation" (LDA) model for document clustering [11]. In this paper, we explicitly show the relationships between the proposed model and the mixed-membership models and discuss the relationships, which we believe will be highly helpful for future elaboration of the statistical method for mutation signature extraction problems.

The R package for the proposed method, **pmsignature** (**p**robabilistic **m**utational signature), is available at `https://github.com/friend1ws/pmsignature`). The core part of the estimation process is implemented in C++ by way of the Rcpp package [12], which enables us to handle millions of somatic mutations from thousands of cancer genomes using standard desktop computers.

Methods

Independence assumption in mutation signatures

The term "mutation signature" is used to describe a characteristic mutational pattern observed in cancer genomes, that is often related to some carcinogens (e.g., significantly frequent C > A mutations in lung cancers with smoking histories). Mathematically, mutation signatures have been characterized as frequencies or probabilities over mutation pattern [6, 8].

Typically, mutation patterns are categorized by 6 substitution patterns (C>A, C>G, C>T, T>A, T>C, T>G, the original base is usually fixed to C or T for removing the redundancy of taking complementary strands), or 96 ($6 \times 4 \times 4$) patterns by considering immediate 5' and 3' flanking bases as well as 6 substitution patterns. Furthermore, taking account of the transcription direction (plus or minus), categorization by 192 mutation patterns is sometimes used [5, 6]. In this paper, we call the factors composing mutation patterns (such as substitution patterns, adjacent bases and transcriptional strand) as "mutation features."

One of the major problems is that total number of mutation patterns exponentially increases as we increase the number of mutation features to take into account. For example, if we want to take account of up to n bases 5' and 3' to the mutated site (which we call $-n$ site and $+n$ site, respectively, in this paper) as well as 6 substitution patterns, the number of parameters becomes $6 \times 4^{2n} - 1$. On the other hand, the mutation rates are mostly below 10 per Mb [5], (where about 30,000 mutations are expected in entire genomic regions). Therefore, considering even ± 2 sites and substitution patterns (1535 parameters) is often difficult because of instability of parameter estimation.

On the other hand, observing the mutation signatures observed in the previous studies, many mutation signatures had clear characteristics such as:

- A certain value of mutation features is often dominant in many known mutation signatures. For example, the base T is highly dominant at the -1 site in the APOBEC signature, and C > T substitution is highly dominant in the ultraviolet signature [5] (see Figure 1(a)).

- Proportional relationships among mutation features are often observed in many signatures as shown in the APOBEC signature, where the frequencies of bases at the +1 site are largely proportional across 3 major substitution patterns C>T, C>G and C>A, and show consistent tendencies (A and T bases are more frequent than C and G bases).

These characteristics imply that the parameter spaces of mutation signatures can be degenerated to lower dimensional spaces in many cases, and we can potentially reduce the number of parameters by imposing additional restrictions.

In this paper, we propose to represent mutation signatures with independent multiple multi-nominal distributions on each mutation feature. By doing this, we can reduce the number of parameters from exponential to linear with respect to the number of mutation features. When we consider up to $\pm n$ sites as well as 6 substitution patterns, the required number of parameters becomes $5 + 6n$. For example, when considering ± 2 sites and substitution patterns, the number of parameters becomes 17, which is far less compared to 1535 in the previous approach. Furthermore, independence assumption enables us to come up with fairly interpretable representation for mutation signatures.

It may be argued that the independence assumption does not conform to real mechanisms of mutational processes because it is highly unlikely that mutagens independently choose each mutation feature such as flanking bases and substitution patterns. However, we would like to present the following justifications for the independence assumption:

1. For representing the transcription factor binding motifs, the position specific weight matrix (PSWM), which is equivalent to independent multiple multinomial distributions, has been quite successfully utilized even though it is never likely that transcription factors independently choose the bases of their binding sites. This is probably because PSWM can capture important characteristics of transcription factor binding sites in spite of the independence assumption, and can be represented in an intuitively interpretable way via sequencing logos [13].

2. As we demonstrate in later sections, the proposed approach can robustly extract most of previously-collected mutation signatures even independence assumption is imposed.

3. Although most mutation signatures show proportional relationships among mutation features, a few signatures seem to violate proportional relationships and cannot be explained by independent models. For example, the pol ϵ mutation signature [5] (see Figure 1(b)) puts strong probability masses on just two pattern TpCpT > TpApT and TpCpG > TpTpG, and the frequencies of substitution patterns and +1 bases are not proportional unlike the APOBEC signature. However, by using multiple mutation signatures, we can represent this phenomenon even with independence assumption.

Mathematical representation of mutation signatures

In this subsection, we show how to mathematically describe mutations collected from genome sequencing studies and mutational signatures. First, let $\Delta^S = \{(t_1, \cdots, t_S) |\ t_s \geq 0 (\forall s = 1, \cdots, S), \sum_{s=1}^{S} t_s = 1\}$ denote S-dimensional simplex, which is used to represent nonnegative vectors summing to 1 throughout the paper.

Suppose each somatic mutation has L mutation features, $\boldsymbol{m} = (m_1, m_2, \cdots, m_L)$, where each m_l can take M_l discrete values. Also, we set $\boldsymbol{M} = (M_1, \cdots, M_L)$. Let $\boldsymbol{x}_{i,j} = (x_{i,j,1}, \cdots, x_{i,j,L}), (i = 1, \cdots, I, j = 1, \cdots, J_i)$ denote the mutation feature vector for the j-th mutation of i-th cancer genome, where $x_{i,j,l} \in \{1, \cdots, M_l\}$, I is the number of available cancer genomes and J_i is the number of mutations in the i-th cancer genome. When taking account of 6 substitution patterns and ± 2 sites, $\boldsymbol{M} = (6, 4, 4, 4, 4)$. See Table 1 for other representation example.

Suppose that there are K mutation signatures, Let $\boldsymbol{f}_{k,l} = (f_{k,l,1}, \cdots, f_{k,l,M_l}) \in \Delta^{M_l}$ denote the multinomial distribution parameter of k-th mutation signature and the l-th mutation feature, then the probabilistic distribution of k-th signature corresponds to $\{\boldsymbol{f}_{k,l}\}_{l=1,\cdots,L}$.

Therefore, each probabilistic mutation signature, in general, consists of multiple multinomial distribution parameters. We give a way of visualizing probabilistic mutation signature (see Figure 2), which is reminiscent of sequencing logos [13].

Non-independent model as a special case of the proposed method

The previous approaches, where each mutation signature is not a set multiple multinomial distributions but one high-dimensional multinomial distribution, can be considered as a special case of the above framework.

Suppose each mutation feature vector (such as substitution patterns, adjacent bases, and so on) is re-labeled as a single-valued mutation feature by some lexicographical order (see Table 1), then each mutation signature becomes one multinomial distribution. We call this representation and mutational process modeling as "full representation," and "full model." In this case, when taking account of 6 substitution patterns and ± 2 sites, $\boldsymbol{M} = (1536)$. On the other hand, the usual representation, where each mutation feature has its corresponding value and individual multinomial distribution, is called as "independent representation," and "independent model."

The full representation model potentially represent complicated mutational processes (e.g., a situation where C > A is frequent at ApCpG sites and C > T is frequent at TpCpA sites) with one signature. However, when many mutation contextual factors are taken into account and the number of free parameters get huge, estimated results tend to be unstable and unreliable. Furthermore, there is a great risk that we over-interpreting the seemingly complex estimated results.

Generative model of mutation features for each cancer genome

Each cancer genome naturally has multiple mutation signatures, because we are living exposed to a variety of carcinogens. Also, the strength of each mutation signature varies among cancer samples, depending on lifestyles, genetic difference and so on. We represent the distribution of mutation signatures for the i-the cancer genome by $\boldsymbol{q}_i = (q_{i,1}, q_{i,2}, \cdots, q_{i,K}) \in \Delta^K, (i = 1, \cdots, I)$.

We adopt a two step model for a generative model of mutations. First, one of the contributing mutation signature is chosen for each mutation depending on the signature distribution parameter of the corresponding cancer genome $\{\boldsymbol{q}_i\}$. Then, according to the probability distribution of the selected mutation signature, mutation features such as substitution patterns and flanking base pairs are generated.

The detailed description of generative process of $\{\boldsymbol{x}_{i,j}\}$ is as follows: For the j-th mutation in the i-th cancer genome,

1. Generate $z_{i,j} \sim \text{Multinomial}(\boldsymbol{q}_m)$, where $z_{i,j} \in \{1, \cdots, K\}$ is the underlying mutation signature causing that mutation.

2. For each $l(= 1, \cdots, L)$, generate the values of each mutation feature $x_{i,j,l} \sim \text{Multinomial}(\boldsymbol{f}_{z_{i,j},l})$.

Relationship with mixed-membership models

The two step generative model described in the previous subsection has close relationships with the mixed-membership models that have been adopted in many other applications, such as document classification and population structure inference problems. In this subsection, we show the relationships between the proposed method and mixed-membership models, slightly abusing notations to contrast the relationships with the proposed method.

In topic models [11, 14], which is a form of mixed-membership models frequently used in document classification problems, each document is assumed to have K different "topics" in varying proportions ($\boldsymbol{q}_i \in \Delta^K$), where each topic is characterized by a word frequency (a multinomial distribution on a set of words W ($\boldsymbol{f}_k \in \Delta^W$). And each word is assumed to be generated by one of K multinomial distributions (topics). The detailed generative process of the j-th word in the i-th document $x_{i,j}$ is:

1. Generate the underlying topic for the j-th word, $z_{i,j} \sim \text{Multinomial}(\boldsymbol{q}_i)$., where $z_{i,j} \in \{1, \cdots, K\}$.

2. Generate $x_{i,j} \sim \text{Multinomial}(\boldsymbol{f}_{z_{i,j}})$, where $x_{i,j} \in \{1, \cdots, W\}$.

Actually, the proposed model in case of $L = 1$ or "full representation" is mostly the same as topic models.

On the other hand, in population structure inference problems [10, 15], each individual is assumed to be an admixture of K ancestries in varying proportions, where each ancestry is characterized by the allele frequency at each SNP locus, and each SNP genotype of an individual are assumed to be generated by the two step model: first, one of the ancestries are chosen by the admixture ratio of ancestries given for each individual, and then the SNP genotype is generated according to the allele frequency of the selected ancestry at that locus. The relationships among the mutation signature model, topic models and population structure models are summarized in the Table 2.

Furthermore, close relationships between mixed-membership models and nonnegative matrix factorization, which has been successfully used in the previous studies for mutational signature problems [4–6], have been pointed out [16]. In fact, the proposed method can be seen as non-negative matrix factorization with additional restrictions. See Supplementary Materials for the detail of the relationships between the proposed approach and nonnegative matrix factorization.

Estimating parameters

The parameters $\{\boldsymbol{f}_{k,l}\}$ and $\{\boldsymbol{q}_i\}$ are not given and have to be estimated from the available mutation data $\{\boldsymbol{x}_{i,j}\}$. On estimating the parameters for mixed-membership models, a number of approaches have been proposed.

After realizing the relationships with mixed-membership models, a number of past parameter estimation techniques proposed for mixed-membership models can be tailored for the proposed model. EM-algorithm (or its variant, called tempered EM algorithm) have been adopted in classical topic models for document classification ([14]) and population structure estimation ([17]). In the population structure estimation problem, [15] has proposed a fast block relaxation scheme using sequential quadratic programming for block updates with a quasi-Newton acceleration of convergence [18], demonstrating a great improvement over EM-algorithm. Similar techniques are used in [19] for the document classification. For the estimation of Bayesian mixed-membership models, (collapsed) Gibbs sampling [10,20] and variational method [11,21,22] have been proposed.

In this paper, we adopt a relatively simple approach that uses EM-algorithm. However, there is a great possibility that we can devise far more efficient approach based on past experiences. Let $g_{i,\boldsymbol{m}}$ denote the number of the i-th sample's mutations with the mutation feature vector $\boldsymbol{m}$.

Introducing the auxiliary variables $\theta_{i,k,\boldsymbol{m}}$, we update these auxiliary variables in the E-step as

$$\theta_{i,k,\boldsymbol{m}} = \frac{q_{i,k} \prod_{l=1}^{L} f_{k,l,m_l}}{\sum_{k'=1}^{K} q_{i,k'} \prod_{l=1}^{L} f_{k',l,m_l}}.$$

Then, in the M-step, we update the parameters $\{f_{k,\boldsymbol{m}}\}$ and $\{q_{i,k}\}$ as

$$f_{k,l,p} = \frac{\sum_{\boldsymbol{m}:m_l=p} g_{i,\boldsymbol{m}} \theta_{i,k,\boldsymbol{m}}}{\sum_{p'} \sum_{\boldsymbol{m}:m_l=p'} g_{i,\boldsymbol{m}} \theta_{i,k,\boldsymbol{m}}},$$

$$q_{i,k} = \frac{\sum_{\boldsymbol{m}} g_{i,\boldsymbol{m}} \theta_{i,k,\boldsymbol{m}}}{\sum_{k'} \sum_{\boldsymbol{m}} g_{i,\boldsymbol{m}} \theta_{i,k',\boldsymbol{m}}}.$$

In addition, we use SQUAREM [23], which is a general framework for accelerating the convergence of any fixed-point iteration such as EM algorithm. Furthermore, to reduce problems with convergence to local minima, we performe EM-algorithm several times (10 times in this paper) changing the initial points, and adopt the estimate with the maximum log-likelihood. For the derivation of EM-algorithm, please see Supplementary Material.

Adding background signatures

There may be a possibility that the intrinsic composition of the genome sequence influences the estimated mutation signatures. For example, the number of observed C > T transitions at CpG sites may increase at promoter regions, just because CpG dinucleotides are more frequent in those regions. In the previous research [8], this background problem was dealt by explicitly incorporating mutation "opportunity" coefficients into the model.

Here, to offset the influences of intrinsic sequence composition, we add background signatures $\{\boldsymbol{f}_{0,\boldsymbol{m}}\} \in \Delta^{M_1 \times \cdots \times M_L}$. For example, when obtaining the background signatures in case of considering substitution patterns and up to ± 2 bases, we first calculate the frequencies of 5-mers where complement sequences are taken when the central bases are A or G, and then the frequencies divided by 3 was give on each mutation feature vectors $\boldsymbol{m}$ considering the alternated bases are equally likely. Since we mainly deal with mutation data from exome sequencing, background signatures were calculated on entire exonic regions.

Estimating standard errors

The standard errors for the parameter estimates are calculated using bootstrap, where somatic mutations are resampled according to the empirical distribution of the original data $\{\boldsymbol{x}_{i,j}\}$ for each cancer genome. For each bootstrap resample, we performed the re-estimation using parameters obtained for the original data as initial points, and calculated sample standard errors of the inferred mutational signatures as estimates of parameter standard errors. In this paper, we performed 100 bootstrap resampling.

Selecting the number of signatures

Determining the number of mutation signatures K is an important and challenging task. One approach is to utilize some statistical information criteria such as AIC [24], BIC [25]. In the population structure problems, for example, the Bayesian deviance [10], and cross-validation [26] have been suggested. One previous study on mutation signature problems [8] utilized BIC. The problem of using these statistical information criteria is that most of them are based on the likelihood, where slight deviations between the specified probabilistic models and the reality sometimes lead to larger number of mutation signatures for compensating those deviations, and produce results with a risk of over-interpretation.

In this paper, instead of utilizing statistical information criteria, we adopt following strategies:

- After calculating the likelihood and standard errors of parameters for a range of K, the value of K is determined at the point where the likelihood is sufficiently high, and the standard errors are sufficiently low [6].
- When, for k_1-th and k_2-th mutation signatures, we could detect strong correlations between the estimated membership parameters for each cancer genome ($(q_{1,k_1}, q_{2,k_1}, \cdots, q_{I,k_1})$ and $(q_{1,k_2}, q_{2,k_2}, \cdots, q_{I,k_2})$), and the two mutation signatures ($\{f_{k_1,\boldsymbol{m}}\}$ and $\{f_{k_2,\boldsymbol{m}}\}$) show similar patterns, then it is likely that an excess amount of K forced to split one mutation signature into two. We stop increasing K before these pairs of mutation signatures are observed.

The strategies listed above are probably not exhaustive, and we should add other practical strategies observing various situations. In addition, it would be nice is we could devise automated and practical approaches for choosing K, which is a possible future challenges.

Results

Experiments on synthetic data

First, we would like to investigate whether the proposed approach can extract "true" mutation signatures or not. Since we can not know the true mutation signatures in real biological data, we resorted to simulation studies.

Here, we generated a set of mutations changing the number of cancer genomes (10, 25, 50, 100), and the number of mutations for each cancer genome (10, 25, 50, 100, 250, 500, 1000). The number of mutation signatures were set to 5 including background mutation ratio. The mutation feature parameters and signature distribution parameters were generated by Dirichlet distribution,

$$f_{k,l,m_l} \sim \mathrm{Dir}(\alpha \mathbf{1}),\ k = 1, \cdots, K,\ l = 1, \cdots, L.$$
$$q_{i,k} \sim \mathrm{Dir}(\gamma \mathbf{1}),\ i = 1, \cdots, I,$$

where the α and γ represent the amounts of dispersion for the mutation features parameters and signature distribution parameters, respectively. When these values are smaller, only a fewer components can have larger probability masses. and the rest will have much smaller masses. When these are large, all the components tend to have evenly-distributed probability masses.

As Figure 3 shows, we could estimate the mutation signatures very accurately overall. (see Supplementary Figure 1 for an example). In most cases, the log-likelihood stopped increasing at five mutation signatures, whereas the standard error of the estimated parameters started increasing past five mutation signatures (see Supplementary Figure 2), indicating that true number of mutation signatures can be recovered by observing the trade-off between likelihood and standard-errors at least in an ideal case.

As expected, as we increase the number of either cancer genomes or mutations, the accuracy of the estimates increased. Also, as we increase the value of α, then the accuracy of the estimated signatures got worse. However, even we increase the value of γ, the accuracy of the estimate signatures did not changed much. Therefore, the dispersion of mutation feature parameters influences the accuracy of mutation signatures more sensitively than signature distribution parameters.

Experiments using cancer genomes from urothelial carcinoma of the upper urinary tract

In this section, we compare the "full model" and the "independent model" by examining mutation signatures obtained by the two approaches, and investigate the robustness by downsampling experiments. The dataset used here is a list of 14717 somatic substitutions collected from the study of 26 urothelial carcinomas of the upper urinary tract (UTUC) [27], where they found a novel mutation signature: T > A substitutions at CpTpG sites with a strong strand specificity caused by aristolochic acids (AA).

We performed the proposed method considering substitution patterns, up to ± 2 sites and strand directions as mutation features. For the full model, we assign one integer for each combination of these features, and thus $L = 1, \boldsymbol{M} = (3072)$. For the independent model, $L = 6, \boldsymbol{M} = (6, 4, 4, 4, 4, 2)$.

First, performing on various number of mutation signatures K (including a background signature) on the independent model, we observed how the detected mutation signatures change (See Supplementary Figure 3). For $K = 2$, a mutation signature which seem to correspond to AA (T > A substitutions at CpTpG sites with strong transcription direction) was observed (Supplementary Figure 3 (a)). For $K = 3$, an additional mutation signature corresponding APOBEC enzyme (C > [AGT] at TpCpN sites) was observed in addition to the AA mutation signature (Figure 4(a), 4(b), Supplementary Figure 3 (b)). For $K = 4$, an additional signature (T > A at NpTpN sites with strong transcription direction) that is somewhat similar to the AA signature was observed (Supplementary Figure 3 (c)). When we checked the correlation of estimated membership parameter for each cancer genome, strong correlation between the

AA signature (CpTpG > CpApG) and that additional AA-like signature could be observed ($R = 0.77$, see Supplementary Figure 5). This additional signature may be just making up for the residual of the AA signature which the original AA signature could not explain due to a slight deviance of the probabilistic model. The values of likelihood continued increasing as the K increase, while the bootstrap-errors started to increase at $K = 5$ (See Supplementary Figure 4). The strong correlation among estimated membership parameters started to be shown at $K = 4$. Considering all these factors together, $K = 3$ seems to be a reasonable choice in terms of the interpretability, and we adopted $K = 3$ in the following.

For the independent model, we could observed the depletion of G base at the -2 site, which is consistent to the previous study [6] and the result in the next subsection. On the other hand, for the full representation model, this tendency was rather mild (Figure 4(c), 4(d), Supplementary Figure 6). Inferred AA mutation signature had no clear characteristics at two bases 5' and 3' to the mutated site compared to the APOBEC mutation signature.

Assuming that the mutation signatures obtained using whole 14717 substitutions as a gold standard, we performed the proposed method on down-sampled data (1%, 2.5%, 5%, 10%, 25%, 50%), and compared obtained signatures with the gold standard. For measuring the deviations of the mutation signature, the cosine similarity was used on the $\prod_{l=1}^{L} M_l$ dimensional vector space $f_{i,\boldsymbol{m}} = \prod_{l=1}^{L} f_{k,l,m_l}$, so that the comparison between the full model and the independent model become possible. Trials for downsampling experiment for each ratio and model were repeated for 100 times.

As the Figure 4(e) and 4(f) shows, the independent representation could recover the original signatures even when, e.g, about 90 % of the data was removed. In this experiment, the AA signature was more robust than the APOBEC signature. We believe that this is because the number of T > A substitutions at GpTpC sites are far more frequent in this dataset. These results indicate that the independent model can give more robust estimates than the full model.

Application to somatic mutation data of 30 cancer types

Finally, we have applied the proposed method for the somatic mutation data of 30 cancer types [5]. For each cancer type, the proposed method was applied separately to see the robustness of the estimated signatures among different cancer types. By changing the number of mutation signatures and calculating values of the log-likelihood and bootstrap errors, we have chosen the number of mutation signatures for each cancer type. The estimated mutation signatures across cancer types were clustered by Frobenius Distance.

The Figure 5, 6 shows the summary of the obtained mutation signatures. In total, 27 mutation signatures were extracted. By comparing the composition of nucleotides and memberships of signatures with the result of the previous study, we could associate many of the detected signatures with known mutational processes.

The signature 1 and 8 (C > A at TpCpT and C > T at TpCpG, respectively) were observed in colorectal and uterine cancers, are associated with deregulated activity of the error-prone polymerase Pol ϵ. The proposed method used two signatures for representing the signature for the Pol ϵ dysfunction, while was represented by one signature in the previous study. Interestingly, signature 1 showed transcription biases whereas signature 8 did not. This phenomena are consistent for both colorectal and uterus cancers (Figures 7(c) and 7(d)). C >A at TpCpT mutations were shown to be enriched in leading strands of replication forks. whereas, there is no mention about the replication strand-specificity of C > T at TpCpG mutations [28]. Although replication strand biases is different from the transcription strand biases, these two biases may have strong correlation, considering the fact that replication origins prefer transcription start sites [29]. Therefore, the transcription strand bias observed in the signature 1 may actually come from replication strand biases.

Furthermore, both the signature 1 and 8 signatures shows strong abundance of T base at the -2 position, and abundance of T bases at two bases 5' and 3' to the mutated site, respectively (Figures 7(c) and 7(d), Supplementary Figure 7(b, c)). In the previous study of Pol ϵ, a nonsense mutation in TP53

at R23X were shown to be enriched in cancers with Pol ϵ defects. In fact, this mutation is C > T at TpTpCpGpA, the most likely patten with the signature 8. This implies that observing ± 2 bases may be helpful for identifying the cause of mutations with higher resolution.

The signature 2 (C > A at [CT]pCpT), observed in low grade gliomas, can be related to the one detected in the same cancer type in the previous study, but slightly different (see Supplementary Figure 8 (a)). This seems to be because the corresponding signature in the previous study shows very complex pattern (C > A at NpCpT or C > T at GpTpN). Actually, the mutation related to this signature is mostly coming from one sample with a very high mutation rate (Supplementary Figure 8 (b,c)), and the signature 2 disappeared when we re-analyze the data removing this sample (Supplementary Figure 9). We suspect that the complex low-grade-glioma specific signature detected in the previous study may be aggregation of multiple signatures, but cannot be decomposed due to the lack of the number of samples having these signatures. Therefore, we need to be more careful for the signature 2 and the corresponding signature in the previous study.

The signature 4 (C > A at CpCpG) observed in kidney clear cell carcinomas, lung adenocarcinomas and melanomas shows remarkably high ratio of base C at the -2 position (see Supplementary Figure 10 (a,b,c)). However, this signature is confide to only a few samples for every cancer type (see Supplementary Figure 10 (d, e, f)). This seems to correspond to the "signatures R2" detected in kidney clear cell carcinomas, lung adenocarcinomas, lung squamous carcinomas and melanomas (see supplementary figures of the previous study [6]) that could not be validated in the previous study, meaning that significant ratios of somatic mutations corresponding to that signature was not estimated to be false positives by re-sequencing or visual inspection of BAM files using genomic viewers. Actually, the pattern of this signature largely matches to that of the putative artifact by oxidation of DNA during acoustic shearing suggested in [**?**], we guess that this signature is the same as what is observed in that study. Although this signature may be from the artifacts, knowing higher resolution of the artifact signature is very helpful to accurately remove the false positive mutations.

The signature 13 (T > [AGT] at TpCpN sites) was observed in 12 cancer types, and surely related to the activity of APOBEC family. The composition of nucleotides in estimated signatures in 12 different cancer types are highly consistent (Figure 7(a)) except for the on in B-cell lymphoma, and almost of of them showed slightly higher abundance of A and T bases and depletion of G base at the -2 position (Figure **??**, Supplementary Figure 7(a)). This fact backs up the results obtained from UCUT data in the previous subsection and the previous study [6], and implies the robustness of the proposed approach. The estimated transcribed strand-specificities varied across cancer types, implying that there is no significant strand-specificity in APOBEC signatures. The Signature 15 and 16 may also be delated to the APOBEC signatures although the estimated forms are slightly different, and are not merged to the cluster of signature 13 in the current criteria.

The signature 3 and 5 (C > A at NpCpN) observed in head-and-neck cancers and three types of lung cancers are probably associated with tobacco smoking. The estimated signature in each cancer type shows higher mutation prevalence on the template strand, which is consistent with the previous study [2, 6]. The signature 6 (C > A at NpCp[AT]) observed in neuroblastomas matches the pattern of that detected in the same cancer type in the previous study. The signature 7 (C > T at NpCpG sites) was observed in 25 out of 30 cancer type, and related to deamination of 5-methyl-cytosine. The signature 9 (C > T at NpCp[CT]) were observed in melanomas and glioblastomas, and are probably associated with a chemotherapy drug, temozolomide. The signature 10 (C >T at [CT]pCpC) were observed in head and neck cancers and melanomas, and probably related to ultraviolet light. Observed consistent strand specificities (Figure 7(e)) matches to the previous results [6]. Also, higher frequencies of T base at the ± 2 position are consistently observed (Figure 7(e), Supplementary Figure 7(d)). The signature 11 (C > T at GpCp[CG]) were observed in small-cell lung cancers and stomach cancers, and seems to be the same with the "signature 15" in the previous study, whose function is still not clear, and the abundance of G base at the ± 2 position is consistently observed (Figure 7(f), Supplementary Figure 7(e)).

The signature 18 (T > C at ApTp[AG]) observed in liver cancer has been shown to be more common in Asian cases than in other ancestries [**?**], though the source of this signature is still not clear. In this signature, we can observe very strong strains specificity implying the evidence of transcription-coupled nucleotide excision repair as shown in [**?**, 6]. The unknown signatures 12 (C > T at [CG]pCp[CT]), 19 (T > C at GpTpN) and 21 (T > [CG] at CpTpT) observed in pilocytic astrocytomas, stomach cancers and oesophagus cancers, respectively, coincides well with the those detected in the same cancer types in the previous study [6]. Although several interesting characteristics were observed at the ± positions for these signatures, their validity are difficult to confirm at this point since these are not reproduced in multiple cancer types.

In summary, the propose method can capture many of the prominent mutation signatures observed in the previous study. Also, the propose method detected several novel characteristics at the ±2 position that is consistent across multiple cancer types, suggesting the existence of more hidden characteristics at the distant position from the mutated site.

Discussion

In this paper, we proposed a novel framework for extracting mutation signatures from a set of somatic point substitutions. We have shown close relationships with mixed membership models, that have been deeply investigated in statistical genetics and machine learning fields, as well as nonnegative matrix factorization. We have demonstrated that the proposed method could capture meaningful characteristics of mutational patterns.

There is a lot of room for improvement by learning from past experiences and knowledge. First, introducing certain prior distributions or penalty terms can lead more efficiency in terms of both accuracy and interpretation, considering the success in machine learning and statistical genetics communities. [30,31]. In the documentation classification problem, adopting determinantal point process priors [32, 33] is demonstrated to be helpful for obtaining more diverse and intuitively interpretable topics, avoiding over fine-tuning for frequent groups of words with multiple similar topics. Second, utilizing other metrics than Euclidean distance and Kullback-Leibler Divergence might be worth investigating. For nonnegative matrix factorization, the metric called β divergence [34, 35], has been successfully utilized in many research fields (reviewed in [36]). Third, for avoiding the problem of determining the number of latent variables, introducing Hierarchical dirichlet processes [37], which is a natural extension of topic models to nonparametric Bayesian frameworks, might be helpful.

In this paper, we have demonstrated that the independent representations could obtain more robust estimates than the full representations. Also, even in the independent representations, more complex mutational process could be represented by utilizing multiple mutation signatures. However, we still do not know about the exact mechanisms of mutational processes and there might be those that should be represented by the full representations or somewhere between the full representations and the independent representations. We should keep exploring appropriate representations of mutation signature with expertise from biology and chemistry.

We have just focused on somatic substitutions in this paper. However, there are a variety of mutations in cancer genomes, such as insertions, deletions, double nucleotides substitutions, structural variations and copy number alterations. Our framework can potentially treat those mutations by considering appropriate mutation features. For example, for deletions, potential mutation features can be the lengths of deletions, adjacent bases, and so on. However, detailed investigation on what mutation features are practical is a future problem.

Acknowledgments

The first author would like to thank to Dr. Daichi Mochihashi for helpful discussion and comments on earlier version on the proposed method. Many of the contents in this paper is deeply influenced by what was going on at Matthew Stephens Laboratory, when the first author stayed at the University of Chicago as a visiting scholar. The first author would like to thank to the members in Matthew Stephens and John Novembre Laboratory, especially Dr. John Novembre, Dr. Jacob Degner for helpful discussion and comments.

References

1. Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. Nature 458: 719–724.

2. Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, et al. (2002) Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene 21: 7435–7451.

3. Pfeifer GP, You YH, Besaratinia A (2005) Mutations induced by ultraviolet light. Mutat Res 571: 19–31.

4. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, et al. (2012) Mutational processes molding the genomes of 21 breast cancers. Cell 149: 979–993.

5. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, et al. (2013) Signatures of mutational processes in human cancer. Nature 500: 415–421.

6. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR (2013) Deciphering signatures of mutational processes operative in human cancer. Cell Rep 3: 246–259.

7. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401: 788–791.

8. Fischer A, Illingworth CJ, Campbell PJ, Mustonen V (2013) EMu: probabilistic inference of mutational processes and their localization in the cancer genome. Genome Biol 14: R39.

9. Krawczak M, Ball EV, Cooper DN (1998) Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet 63: 474–488.

10. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

11. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3: 993–1022.

12. Eddelbuettel D, François R, Allaire J, Chambers J, Bates D, et al. (2011) Rcpp: Seamless r and c++ integration. Journal of Statistical Software 40: 1–18.

13. Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18: 6097–6100.

14. Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: ACM, SIGIR '99, pp. 50–57. doi:10.1145/312624.312649. URL `http://doi.acm.org/10.1145/312624.312649`.

15. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19: 1655–1664.

16. Ding C, Li T, Peng W (2008) On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. Computational Statistics & Data Analysis 52: 3913–3927.

17. Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: analytical and study design considerations. Genetic epidemiology 28: 289–301.

18. Zhou H, Alexander D, Lange K (2011) A quasi-newton acceleration for high-dimensional optimization algorithms. Statistics and computing 21: 261–273.

19. Taddy MA (2011) On estimation and selection for topic models. arXiv preprint arXiv:11094518 .

20. Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci USA 101 Suppl 1: 5228–5235.

21. Teh YW, Newman D, Welling M (2006) A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In: Advances in neural information processing systems. pp. 1353–1360.

22. Raj A, Stephens M, Pritchard JK (2014) Variational inference of population structure in large snp datasets. Genetics : genetics–114.

23. Varadhan R, Roland C (2008) Simple and globally convergent methods for accelerating the convergence of any em algorithm. Scandinavian Journal of Statistics 35: 335–353.

24. Akaike H (1974) A new look at the statistical model identification. Automatic Control, IEEE Transactions on 19: 716–723.

25. Schwarz G, et al. (1978) Estimating the dimension of a model. The annals of statistics 6: 461–464.

26. Alexander DH, Lange K (2011) Enhancements to the admixture algorithm for individual ancestry estimation. BMC bioinformatics 12: 246.

27. Hoang ML, Chen CH, Sidorenko VS, He J, Dickman KG, et al. (2013) Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. Sci Transl Med 5: 197ra102.

28. Shinbrot E, Henninger EE, Weinhold N, Covington KR, Goksenin AY, et al. (2014) Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. Genome Res 24: 1740–1750.

29. Dellino GI, Cittaro D, Piccioni R, Luzi L, Banfi S, et al. (2013) Genome-wide mapping of human DNA-replication origins: levels of transcription at ORC1 sites regulate origin selection and replication timing. Genome Res 23: 1–11.

30. Hoyer PO (2004) Non-negative matrix factorization with sparseness constraints. The Journal of Machine Learning Research 5: 1457–1469.

31. Engelhardt BE, Stephens M (2010) Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. PLoS genetics 6: e1001117.

32. Kulesza A, Taskar B (2012) Determinantal point processes for machine learning. arXiv preprint arXiv:12076083 .

33. Kwok JT, Adams RP (2012) Priors for diversity in generative latent variable models. In: Advances in Neural Information Processing Systems. pp. 2996–3004.

34. Basu A, Harris IR, Hjort NL, Jones M (1998) Robust and efficient estimation by minimising a density power divergence. Biometrika 85: 549–559.

35. Eguchi S, Kano Y (2001) Robustifying maximum likelihood estimation. Tokyo Institute of Statistical Mathematics, Tokyo, Japan, Tech Rep .

36. Févotte C, Idier J (2011) Algorithms for nonnegative matrix factorization with the β-divergence. Neural Computation 23: 2421–2456.

37. Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical dirichlet processes. Journal of the american statistical association 101.

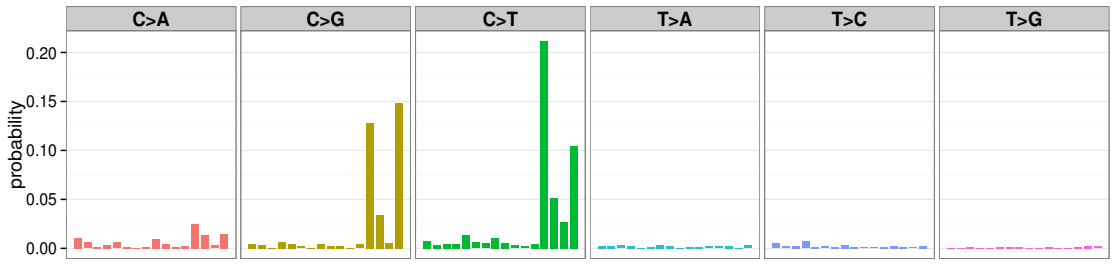
Tables

Table 1. Example of representation for mutation patterns (substitution patterns and one 5' and 3' bases) In the independent representation, the elements of vector show substitution patterns, 5' adjacent bases and 3' adjacent bases, respectively. For the substitution pattens, 1 to 6 values are assigned to C>A, C>G, C>T, T>A, T>C and T>G in this order. For 5' and 3' adjacent bases, 1to 4 values are assigned to A, C, G and T. Note that the original base is fixed to C or T to remove the redundancy of complement sequences.

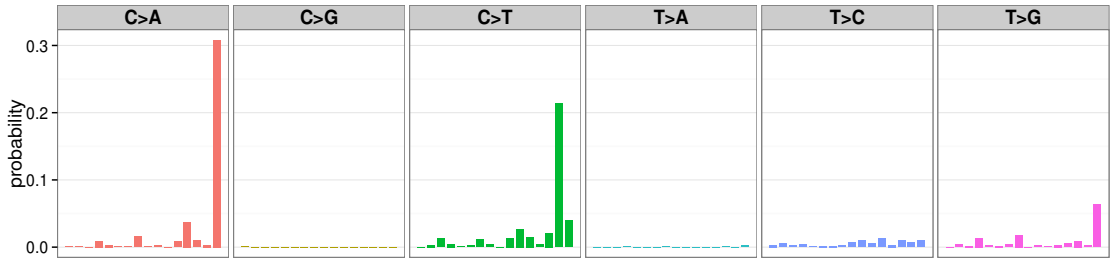
mutation pattern	full representation	independent representation
L	1	3
$\boldsymbol{M}$	(96)	(6, 4, 4)
ApCpA → ApCpA	(1)	(1, 1, 1)
ApCpC → ApApC	(2)	(1, 1, 2)
ApCpG → ApApG	(3)	(1, 1, 3)
ApCpT → ApApT	(4)	(1, 1, 4)
CpCpA → CpApA	(5)	(1, 2, 1)
...	...	...
ApCpA → ApGpA	(17)	(2, 1, 1)
...	...	...
TpTpT → TpGpT	(96)	(6, 4, 4)

Table 2. Relationships among mutation signature model, topic models, and population structure models.

problem	$\boldsymbol{x}_{i,j}$	$\boldsymbol{f}_k$	$\boldsymbol{q}_i$
mutation signature model	the j-th mutation in the i-th cancer genome	the feature dist. for the k-th signature	the signature dist. for the i-th cancer genome
topic model	the j-th word in the i-th document	the word dist. for the k-th topic	the topic dist. for the i-th document
population structure model	the j-th locus genotype of the individual i	the allele freq. for the ancestry k	the admixture dist. for the individual i

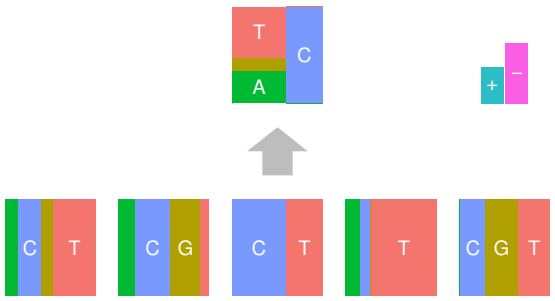


(a) APOBEC signature in the previous study



(b) POLE signature in the previous study

Figure 1. The APOBEC and POLE signatures extracted in the previous study (Signature 2 and 10 in [5], respectively). The barplots are divided by 6 substitution patterns. In each division, 16 bars show joint probabilities of 16 combinations of the immediate 5' and 3' bases (ApNpA, ApNpC, ApNpG, ApNpT, CpNpA, $\cdots$, TpTpT). (a) The strong intensities of the last four bars for the three substitution patterns with original bases C indicates that the immediate 5' base is mostly confined to T in this signature. Also, the frequency of the immediate 3' bases for the mutation patterns at TpCpN sites are mostly proportional across three major substitution patterns (C > A, C > G and C > T). (b) The strong intensities on TpCpT > TpApT and TpCpG > TpTpG are observed. This is a little complex in the sense that the immediate 3' base and the substitution pattern are not independent.



(a) substitution pattern

C>A	C>G	C>T	T>A	T>C	T>G
0.201	0.080	0.314	0.000	0.405	0.000

(b) adjacent bases

position	A	C	G	T
-2	0.146	0.257	0.128	0.469
-1	0.183	0.389	0.331	0.097
+1	0.161	0.107	0.013	0.719
+2	0.019	0.265	0.366	0.350

(c) strand direction

plus strand	minus strand
0.375	0.625

Figure 2. An example of a mutation signature and its visualization in the proposed approach. Here, mutation features (substitution patterns, two 5' and 3' bases and strand direction) are assumed to be independent ($L = 6$, $\boldsymbol{M} = (6, 4, 4, 4, 4, 2)$). In the bottom, the size of each box represents the frequencies of bases (A, C, G and T) at the flanking sites. In the top, the height of each box represents the conditional frequencies of mutated bases for each original base (C and T). In the upper right, the height of the + box represents the frequencies of mutations in the coding strand (or the plus strand, the sense strand and the untranscribed strand) whose nucleotide sequences directly corresponds to mRNA, whereas the height of − box represents those in the template strand (or the minus strand, the antisense strand, the transcribed strand and the noncoding strand) whose sequences are copied during the synthesis of mRNA.

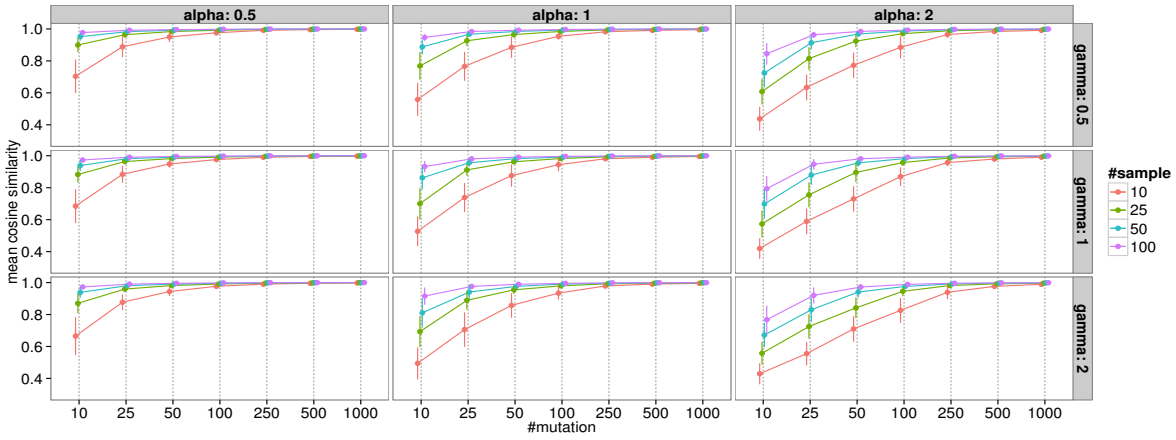
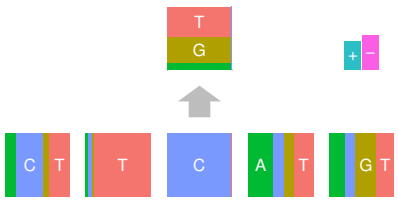
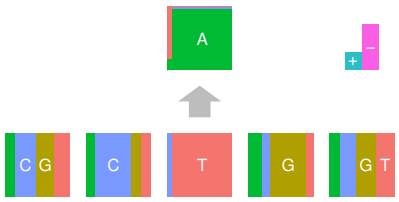


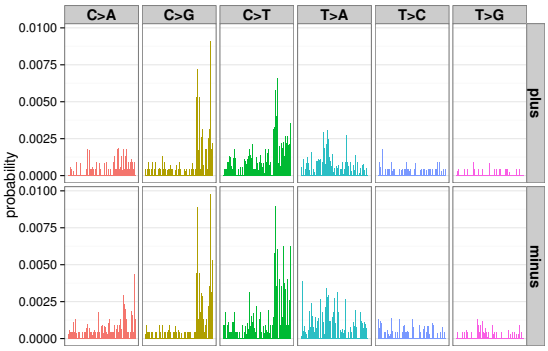
Figure 3. The accuracy of the proposed approach for the simulated data when changing the number of samples, mutations, and the amounts of dispersion parameters (α and γ) for the mutation features and signature distribution parameters.



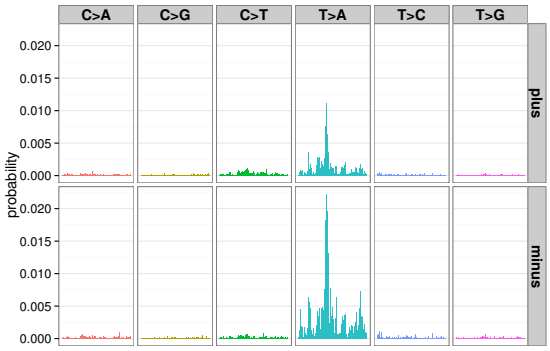
(a) APOBEC signature for the independent model



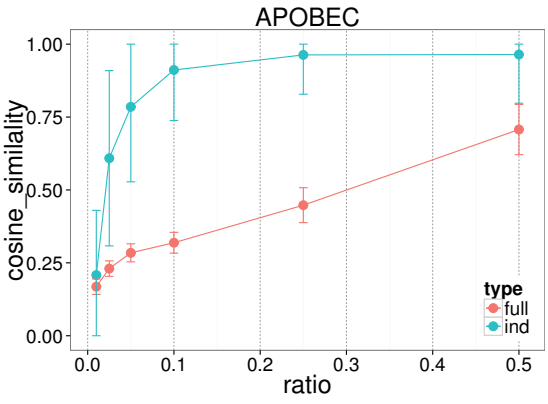
(b) AA signature for the independent model



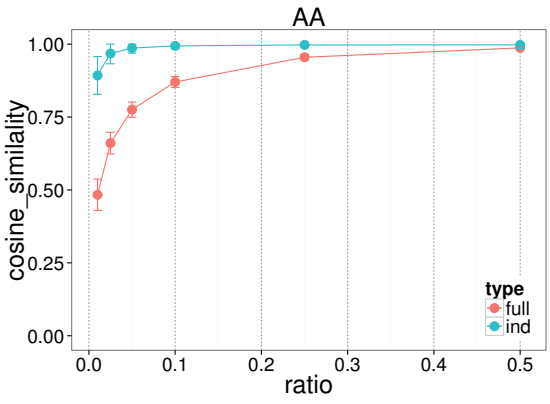
(c) APOBEC signature for the full model



(d) AA signature for the full model



(e) APOBEC signature stability



(f) AA signature stability

Figure 4. The mutation signatures for the UTUC data, and the results of downsampling experiments. 3072 elements in the full model mutation signatures were shown divided by 6 substitution patterns and strand directions.

Figure 5. The summary of mutation signatures obtained in a reanalysis of the data of the previous study [5] using the proposed method, where the substitution patterns and two 5' and 3' bases from the mutated sites are taken into account as mutation features.

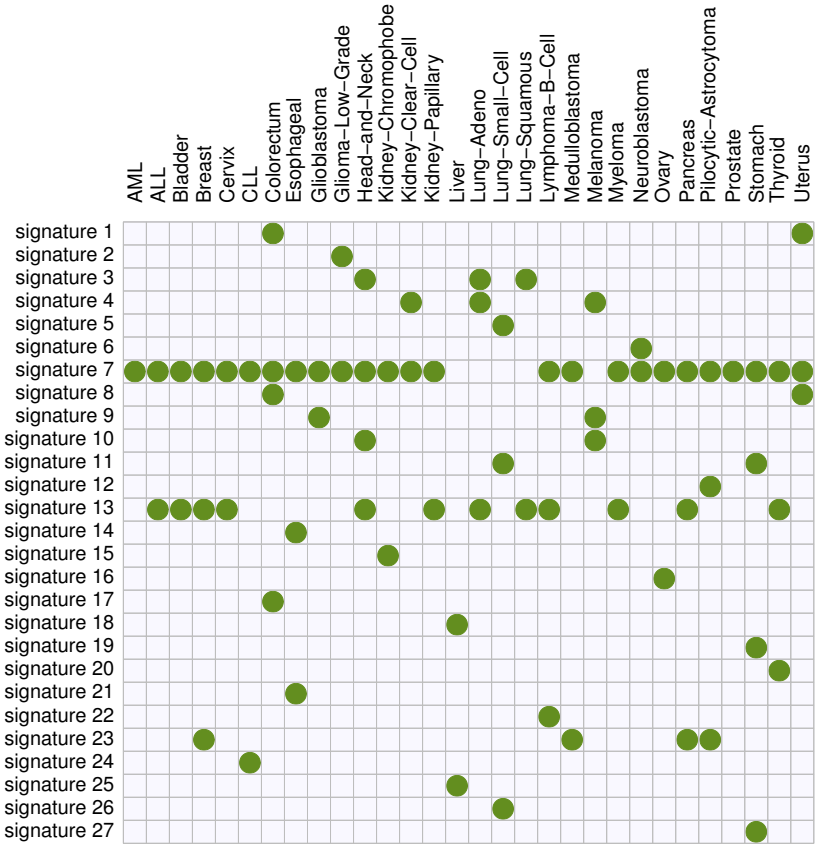
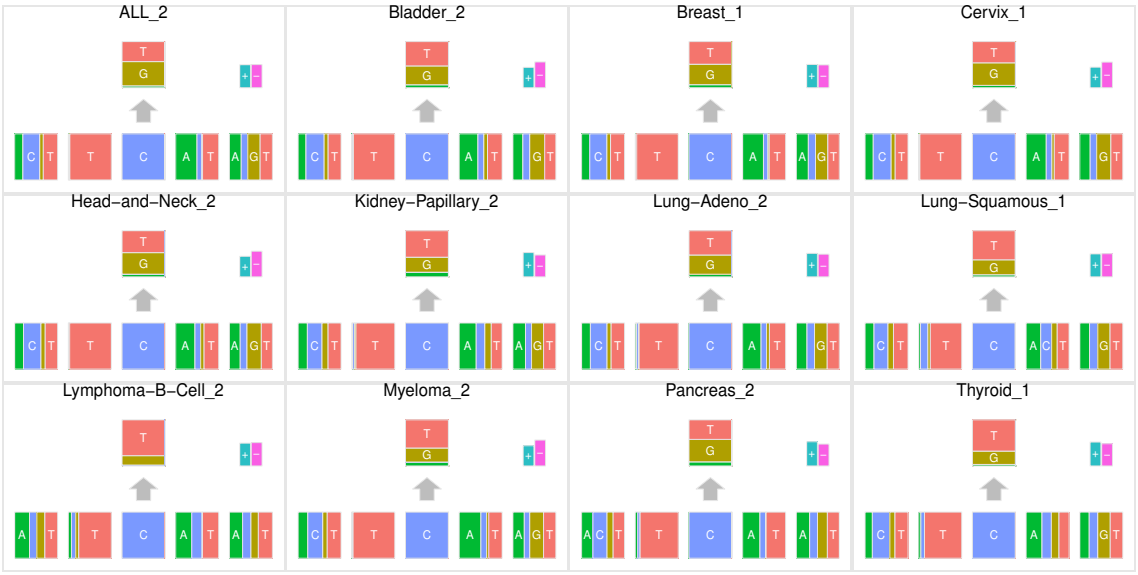
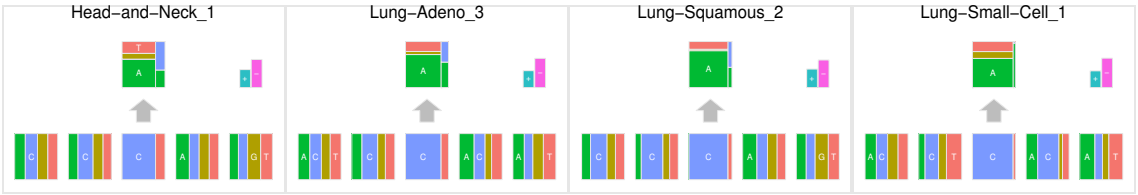


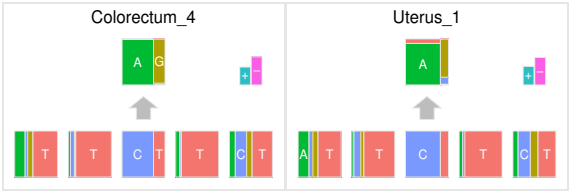
Figure 6. The summary of membership of each mutation signature across 30 cancer types obtained using the proposed method.



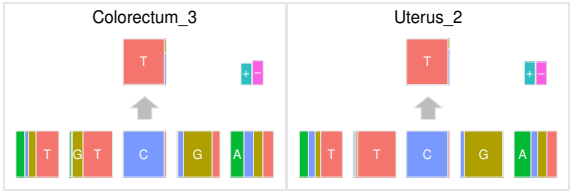
(a) APOBEC signatures obtained in each cancer type



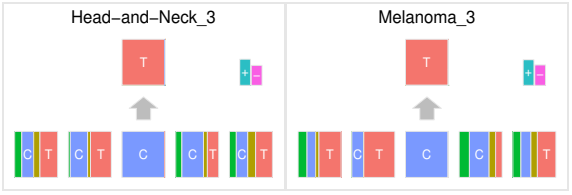
(b) Smoking signature in each cancer type



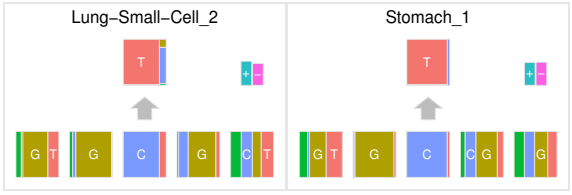
(c) POLE1 signatures in each cancer type



(d) POLE2 signature in each cancer type



(e) UV signature in each cancer type



(f) Unknown signature obtained in lung small cell carcinomas and stomach cancers

Figure 7. (a) APOBEC signatures obtained in 11 cancer types, (b, c, d, e) Several signatures having prominent characteristics at 5' or 3' to the mutated sites, (f) the frequencies of bases at two 5' to the mutated site.