



Genome analysis

# Discovering novel mutation signatures by latent Dirichlet allocation with variational Bayes inference

Taro Matsutani<sup>1,2</sup>, Yuki Ueno<sup>1,2</sup>, Tsukasa Fukunaga <sup>1,3</sup> and Michiaki Hamada <sup>1,2,4,5,\*</sup>

<sup>1</sup>Department of Electrical Engineering and Bioscience, Faculty of Science and Engineering, Waseda University, Tokyo, Japan, <sup>2</sup>AIST-Waseda University Computational Bio Big-Data Open Innovation Laboratory (CBBDOIL), Tokyo, Japan, <sup>3</sup>Department of Computer Science, Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan, <sup>4</sup>Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan and <sup>5</sup>Graduate School of Medicine, Nippon Medical School, Tokyo, Japan

\*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on June 23, 2018; revised on April 3, 2019; editorial decision on April 5, 2019; accepted on April 10, 2019

## Abstract

**Motivation:** A cancer genome includes many mutations derived from various mutagens and mutational processes, leading to specific mutation patterns. It is known that each mutational process leads to characteristic mutations, and when a mutational process has preferences for mutations, this situation is called a ‘mutation signature.’ Identification of mutation signatures is an important task for elucidation of carcinogenic mechanisms. In previous studies, analyses with statistical approaches (e.g. non-negative matrix factorization and latent Dirichlet allocation) revealed a number of mutation signatures. Nonetheless, strictly speaking, these existing approaches employ an ad hoc method or incorrect approximation to estimate the number of mutation signatures, and the whole picture of mutation signatures is unclear.

**Results:** In this study, we present a novel method for estimating the number of mutation signatures—latent Dirichlet allocation with variational Bayes inference (VB-LDA)—where variational lower bounds are utilized for finding a plausible number of mutation patterns. In addition, we performed cluster analyses for estimated mutation signatures to extract novel mutation signatures that appear in multiple primary lesions. In a simulation with artificial data, we confirmed that our method estimated the correct number of mutation signatures. Furthermore, applying our method in combination with clustering procedures for real mutation data revealed many interesting mutation signatures that have not been previously reported.

**Availability and implementation:** All the predicted mutation signatures with clustering results are freely available at <http://www.f.waseda.jp/mhamada/MS/index.html>. All the C++ source code and python scripts utilized in this study can be downloaded on the Internet ([https://github.com/qkirikigaku/MS\\_LDA](https://github.com/qkirikigaku/MS_LDA)).

**Contact:** [mhamada@waseda.jp](mailto:mhamada@waseda.jp)

**Supplementary information:** [Supplementary data](#) are available at *Bioinformatics* online.

## 1 Introduction

Cancer cells carry a number of somatic mutations, which are roughly subdivided into two categories: (i) driver mutations that contribute to the proliferation of cancer genomes and (ii) passenger mutations that do not contribute to the proliferation (Stratton *et al.*, 2009). In the past research, hypothesis-driven studies whose target is driver mutations have mainly been conducted (Smalheiser, 2002). On the contrary, due to the development of next-generation sequencers (NGS), large-scale data on cancer genomes are rapidly accumulated worldwide (Forbes *et al.*, 2015; Tomczak *et al.*, 2015). Accordingly, because many studies that are focused on passenger mutations have been conducted (Rubin and Green, 2009; Wong *et al.*, 2011), passenger mutations are also considered important for understanding carcinogenic mechanisms (Alexandrov *et al.*, 2013a; Barba *et al.*, 2014; Greenman *et al.*, 2007).

Regardless of whether it is a driver mutation or passenger mutation, each somatic mutation has its own cause. For instance, it is known that genome sequences of lung cancer patients with a smoking habit have many typical substitutions, cytosine (C) to adenine (A), in their tumor suppressor genes, such as TP53 (Toyooka *et al.*, 2003). In this case, we call the smoking habit a ‘mutational process’ (causes mutations), and a ‘mutation signature’ is defined as a preference for mutations (mutational distribution, e.g. C to A substitutions frequently occur) corresponding to the mutational process. Every mutational process leads to a specific mutation signature, and accumulation of mutations from birth up to now is considered a result of a combination of some mutation signatures (Stratton, 2011). Hence, clarification of mutation signatures is expected to reveal carcinogenic mechanisms and may serve as biomarkers for early diagnosis (Harris, 2013; Temko *et al.*, 2018; Wagener *et al.*, 2015). Note that Zou *et al.* (2018) recently reproduced mutation signatures with a knockout of individual genes in vitro using the CRISPR-Cas9 technology; this result supports the existence of mutation signatures.

There exist many studies on prediction of mutation signatures by means of cancer mutation catalogs. Nik-Zainal *et al.* (2012, 2016) and Alexandrov *et al.* (2013a, 2015) used non-negative matrix factorization (NMF) (Lee and Seung, 2001) for estimating mutation signatures and revealed 30 mutation signatures in various cancer types (<http://cancer.sanger.ac.uk/cosmic>). The ‘signeR’ also utilized NMF (in combination with the empirical Bayesian approach) for signature discovery (Rosales *et al.*, 2017). On the other hand, Shiraishi *et al.* (2015) used the topic model [in particular, latent Dirichlet allocation (LDA) (Blei *et al.*, 2003)] for modeling and clarifying mutation signatures. In comparison with NMF, topic models assume probabilistic structures of mutations behind samples; this approach is expected to improve generalization performance (Hofmann, 1999).

Although those methods are effective at estimating mutation signatures, they involve ad hoc approaches to prediction of the variety (number) of mutation signatures; this arrangement prevents clarification of the whole picture of mutation signatures. In addition, during a search for a new signature, model selection is an important basis for deciding whether a signature is new. A notable approach to determining the number of mutation signatures is to employ the EMu method (Fischer *et al.*, 2013) where the Bayesian information criterion (BIC) is utilized for model selection. However, BIC is mathematically suitable only for probabilistic models whose Fisher information matrix is regular because the posterior distribution of parameters can be Laplace approximated in the statistical regular model where the central limit theorem holds. Thus, BIC should not

be used in mixed models including latent variables such as LDA (Yamazaki and Watanabe, 2005). Note that signeR described in the previous paragraph also utilized BIC for model selection. To address these issues, we propose a novel method for prediction of mutation signatures and select a plausible model for LDA with variational Bayes (VB) inference. We confirmed that our method is sufficiently accurate in a numerical simulation. In experiments with real datasets, we introduced clustering analyses for estimated mutation signatures in addition to known signatures to discover reliable signatures present in several cancer types. As a result, we found several interesting mutation signatures that could be novel.

## 2 Materials and methods

### 2.1 Representation of mutations in cancer genomes

Due to the recent advances of next-generation sequencers, the number of known cancer genomes is rapidly growing in the world. With a specific cancer genome, we obtain a set of mutations (e.g. a substitution from A to C, denoted by [A > C]) present in the genome. Here, the variety of mutations is defined by a mutation dictionary (denoted by  $\mathcal{M}$ ), and we utilize four mutation dictionaries ( $\mathcal{M}_1$ ,  $\mathcal{M}_2$ ,  $\mathcal{M}_3$  and  $\mathcal{M}_4$ ) according to the purpose of analyses as described below.

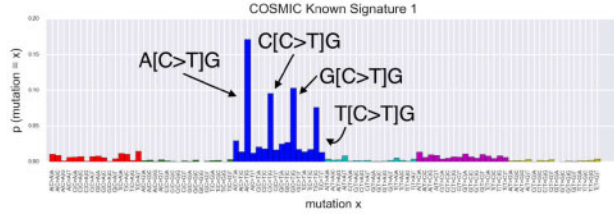
#### 2.1.1 Mutation dictionaries without insertions and deletions (indels)

For a single-base substitution, there are 4 types of bases (A, C, G and T) before the substitution, and 3 types of bases (other than the original base) after the substitution, thus leading to 12 types of mutations. Nonetheless, DNA forms base pairs within the double-stranded structure, and it is impossible to determine the strand on which a mutation occurred using only the observed mutation catalog. For example, whenever the [C > T] substitution happens, substitution [G > A] takes place simultaneously on the complementary strand, and we cannot distinguish these two substitutions. After removal of this redundancy, there are only six kinds of single-nucleotide substitutions (among the 12 types). In this study, we used  $\mathcal{M}_0 := \{[C > A], [C > G], [C > T], [T > A], [T > C], [T > G]\}$ , which is the simplest mutation dictionary (Alexandrov *et al.*, 2013a). Furthermore, Alexandrov *et al.* (2013a) suggests that the bases adjacent to the base where a substitution occurred (called a mutation context) are important; therefore, we include the 5' and 3' adjacent bases into the information on mutations. For instance, it is known that the cytosine that is adjacent to G on the 3' side tends to get methylated, and a methylated cytosine tends to undergo deamination and is prone to change to thymine (Pfeifer, 2006), leading to a specific mutation pattern: 5'-X[C > T]G-3' (X = A, G, C and T), Signature 1 in the COSMIC database (Fig. 1). Motivated by this observation, Alexandrov *et al.* (2013a) introduced mutation dictionary  $\mathcal{M}_1$  for single substitutions and a mutation context for adjacent bases:

$$\mathcal{M}_1 := \{5'-XmY-3'|X, Y \in \{A, C, G, T\}, m \in \mathcal{M}_0\}.$$

Clearly  $|\mathcal{M}_1| = 4 \times 6 \times 4 = 96$  holds.

Additionally, we introduce another mutation dictionary,  $\mathcal{M}_2$ , in which not only adjacent bases but also bases up to 2 units away upstream and downstream from a substitution (e.g. 5'-AT[C > T]GC-3') are considered a mutation context as follows.



**Fig. 1.** A known mutation signature in the COSMIC database. This signature (Signature 1) was taken from the COSMIC database (<http://cancer.sanger.ac.uk/cosmic>), whose mutational process is considered the deamination reaction of methylated cytosine. The horizontal and vertical axes show types of mutations and their probabilities, respectively. The four arrows indicate the peaks at A[C>T]G, C[C>T]G, G[C>T]G and T[C>T]G, suggesting that methylated cytosine tends to become thymine if the 3' adjacent base is guanine

$$\mathcal{M}_2 := \{5'-XYmZW-3' | X, Y, Z, W \in \{A, C, G, T\}, m \in \mathcal{M}_0\}$$

where  $|\mathcal{M}_2| = 4 \times 4 \times 6 \times 4 \times 4 = 1536$  holds. This mutation dictionary is also used by [Alexandrov et al. \(2013a\)](#) and [Shiraishi et al. \(2015\)](#).

### 2.1.2 Mutation dictionaries with indels

The mutation dictionaries in the previous section include only substitutions. It is known that somatic mutations often include many insertions and deletions (indels), some of which involve several tens of bases. We therefore introduce mutation dictionaries including indels. To avoid sparsity of expressions, we consider big ( $\geq 10$  bps) and small ( $< 10$  bps) indels:  $I := \{\text{bigindel}, \text{smallindel}\}$ . As in Section 2.1.1, we consider the mutation context around indels. When we focus on the bases which are adjacent to indels, the combination of bases amounts to  $4 \times 4 = 16$  types. Because the 12 combinations other than 4 combinations ( $[5'-\text{ApXpT-3}']$ ,  $[5'-\text{CpXpG-3}']$ ,  $[5'-\text{GpXpC-3}']$  and  $[5'-\text{TpXpA-3}']$ , which are palindromes with the complementary strand) are considered duplicates, actual mutation contexts consist of 10 patterns. We therefore define  $\mathcal{M}_I$  as the mutation dictionary for a mutation context with respect to indels:  $\mathcal{M}_I := \{AmT, CmG, GmC, TmA, AmC, CmT, TmG, GmA, AmA, CmC | m \in I\}$ . In this study, we introduce the mutation dictionaries with indels,  $\mathcal{M}_3$  (including both substitutions with a simple mutation context and indels) and  $\mathcal{M}_4$  (including both substitutions with a detailed mutation context and indels), which are formally defined as

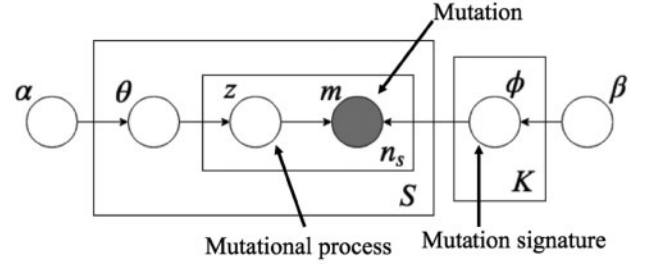
$$\mathcal{M}_3 := \mathcal{M}_1 \cup \mathcal{M}_I \text{ and } \mathcal{M}_4 := \mathcal{M}_2 \cup \mathcal{M}_I$$

respectively. Note that  $|\mathcal{M}_3| = |\mathcal{M}_1| + |\mathcal{M}_I| = 96 + 2 \times 10 = 116$  and  $|\mathcal{M}_4| = |\mathcal{M}_2| + |\mathcal{M}_I| = 1536 + 20 = 1556$ .

## 2.2 LDA for modeling of mutation signatures

By means of one of the mutation dictionaries in the previous section, a mutation catalog in a specific cancer genome is transformed into a set of mutations (defined by the mutation dictionary). We modeled these mutations using a generative probabilistic model called LDA.

LDA was initially proposed as a generative probabilistic model for documents ([Blei et al., 2003](#)) that was successfully applied to some bioinformatics problems, e.g. by [Liu et al. \(2010\)](#) and [Flaherty et al. \(2005\)](#). In LDA, each word in a document is generated from a latent variable called a topic, which has a specific word distribution. In this study, words and topics in LDA correspond to mutations and mutational processes, respectively ([Fig. 2](#)).



**Fig. 2.** A graphical model of LDA for modeling mutation signatures.  $m$  is an observed variable for mutations with a given mutation dictionary,  $\mathcal{M}$  and  $z$  is a latent (hidden) variable for mutational processes. The  $i$ th mutation in the  $s$ th sample,  $m_{s,i} \in \mathcal{M}$ , is generated by a categorical distribution with parameters  $\phi_{z_{s,i}}$ , where  $z_{s,i} \in \{1, 2, \dots, K\}$  corresponds to its mutational process that is generated by a categorical distribution with parameter  $\theta_s$ ;  $\theta$  and  $\phi$  are generated from Dirichlet distributions with hyperparameters  $\alpha$  and  $\beta$ , respectively. In this model,  $\phi_z$  represents a preference for mutations (in  $\mathcal{M}$ ) of mutational process  $z$ , and this arrangement corresponds to a mutation signature.  $S$  and  $K$  indicate the numbers of samples and mutation signatures, respectively, and  $n_s$  is the number of mutations in the  $s$ th sample. In this study, not only parameters  $\alpha$ ,  $\beta$ ,  $\theta$  and  $\phi$  but also  $K$  are estimated from the observed mutations

In this article, we employ the following notations:

- $m_{s,i}$  is the  $i$ th mutation of the  $s$ th sample.
- $S$  is the number of samples;  $s$  ( $1 \leq s \leq S$ ) is an index for a sample.
- $n_s$  is the number of mutations in the  $s$ th sample;
- $V$  is the total number of mutations in mutation dictionary  $\mathcal{M}$ ;  $v$  means the  $v$ th type of mutation ( $1 \leq v \leq V$ ).
- $K$  is the number of mutational processes;  $k$  means the  $k$ th process ( $1 \leq k \leq K$ ).
- $\theta_s = \{\theta_{s,k}\}_{k=1}^K$  is the parameter of the categorical distribution of mutational processes for each sample  $s$ , where  $\theta_{s,k}$  represents the activity of the  $k$ th process in the  $s$ th sample, which is called signature activity.
- $\phi_k = \{\phi_{k,v}\}_{v=1}^V$  is the parameter of the categorical distribution of mutations for each signature  $k$ , where  $\phi_{k,v}$  denotes the proportion of the  $v$ th mutation type for the  $k$ th signature.

Then, we introduce a probabilistic model for the observed mutations  $\{m_{s,i}\}$  with latent variable  $z_{s,i} \in \{1, \dots, K\}$  representing the mutational process of  $m_{s,i}$ . Specifically, the generative model for mutations is expressed as follows:

$$\theta_s \sim \text{Dir}(\alpha), \phi_k \sim \text{Dir}(\beta) \quad (1)$$

$$z_{s,i} \sim \text{Cat}(\theta_s), m_{s,i} \sim \text{Cat}(\phi_{z_{s,i}}) \quad (2)$$

where  $\text{Cat}(\theta)$  and  $\text{Dir}(\alpha)$  represent categorical and Dirichlet distributions with hyperparameters  $\theta$  and  $\alpha$ , respectively. We emphasize that  $\phi_k$ , which represents a mutational signature, is in one-to-one correspondence with the  $k$ th mutation process.

## 2.3 Learning LDA with variational Bayes inference

Either Gibbs sampling or variational Bayes (VB) is frequently used for learning parameters in LDA ([Blei et al., 2003](#)). In this study, we employ VB because the evaluation function of VB (called variational lower bound: VLB) is applicable to model selection (i.e. estimating the number of latent variables). The detailed method is presented in the Supplementary Section S1. In VB, we try to minimize the KL divergence with the true distribution by learning so as to maximize

VLB calculated from the joint distribution without calculating the posterior distribution. By avoiding the calculation of the posterior distribution, it is possible to solve the issue of non-regularity in singular models such as LDA.

In our computational analyses, we updated parameters until the difference between a current VLB value and the previous value by one iteration becomes smaller than  $10^{-5}$  or the number of iterations becomes larger than 1000. To avoid the local minimum, the initial values of parameters were reallocated 50 times for each  $K$ , and the estimated values of the parameters with the highest VLB value were adopted as representative parameters of  $K$ . It should be emphasized that VB inference enables us to select the optimal number of latent variables automatically by taking the highest value of VLB for each  $K$  (Corduneanu and Bishop, 2001), which is a complementary approach to existing approaches such as EMu.

## 2.4 A comparison between predicted and known mutation signatures

After training VB-LDA for a set of mutations with given mutation dictionary  $\mathcal{M}$ , we obtained the number of mutation signatures ( $K$ ) and parameters  $\phi = \{\phi_k\}_{k=1}^K$ , where  $\phi_k$  denotes a probability vector in  $\mathcal{M}$  for mutation signatures  $k$ . On the other hand, 30 mutation signatures are obtained from the COSMIC database (<https://cancer.sanger.ac.uk/cosmic>), and the probability vector of the  $l$ th known signature is denoted by  $\psi_l^{\text{known}}$ . If we choose the mutation dictionary  $\mathcal{M}_1$  in our method, then the predicted signatures are directly comparable to the known signatures. In this study, cosine distance is used for the comparison because it can capture characteristic peaks in a mutation distribution.

For mutation dictionaries  $\mathcal{M}_2$ ,  $\mathcal{M}_3$  and  $\mathcal{M}_4$ , we introduce a probability distribution whose vocabulary is the same as  $\mathcal{M}_1$  as

$$p_2(XmY) := \sum_{Z, W \in \{A, G, C, T\}} p_{\mathcal{M}_2}(ZXmYW)$$

$$p_3(XmY) := \frac{p_{\mathcal{M}_3}(XmY)}{\sum_{X', Y' \in \{A, G, C, T\}} \sum_{m \in \mathcal{M}_0} p_{\mathcal{M}_3}(X'mY')}$$

$$p_4(XmY) := \frac{\sum_{Z, W \in \{A, G, C, T\}} p_{\mathcal{M}_4}(ZXmYW)}{\sum_{X', Y', Z, W \in \{A, G, C, T\}} \sum_{m \in \mathcal{M}_0} p_{\mathcal{M}_4}(X'mY'W)}$$

respectively, for  $X, Y \in \{A, C, G, T\}$ ,  $m \in \mathcal{M}_0$ . Here,  $p_{\mathcal{M}_d}$  denotes the estimated mutational distribution in  $\mathcal{M}_d$  ( $d=2, 3, 4$ ). Given that  $p_2$ ,  $p_3$  and  $p_4$  clearly provide probability vectors for mutation dictionary  $\mathcal{M}_1$ , they are used in comparison with the known COSMIC signatures.

## 3 Results and discussion

### 3.1 Results on simulated mutation datasets

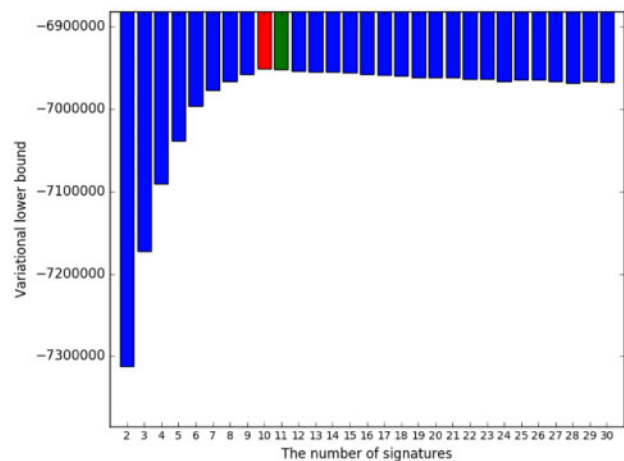
To confirm the usefulness of our proposed model, we conducted experiments with a simulated dataset. In these analyses, we simulated mutation data according to the LDA model, where  $\phi_k$  ( $k=1, 2, \dots, 10$ ) is taken from COSMIC Known Signature 1, 2,  $\dots$ , 10. We generated various simulated data by changing total sample number  $S$ , the number of mutations for one sample  $n_s$ , and hyperparameter  $\alpha_k$  to determine under what conditions the model can predict original signatures correctly. We tested  $K$  from 2 to 30, and the optimal  $K$  was estimated by VB inference (see Section 2.3).

Table 1 shows the results of simulation, and Figure 3 and Supplementary Figure S1 present the VLB for  $K=2, 3, \dots, 30$  and

**Table 1.** Simulation results for various parameters of data generation

# samples: $S$	20	100	1000	5000
# predicted signatures	9	10	10	10
Ave. Cos. dist. (to COSMIC Signatures)	0.0614	0.0096	0.0138	0.0050
Ave. Cos. dist. (to predicted signatures)	0.0473	0.0096	0.0138	0.0050
# mutations in sth sample: $n_s$	100	200	400	2000
# predicted signatures	6	7	10	10
Ave. Cos. dist. (to COSMIC Signatures)	0.0276	0.0262	0.0462	0.0138
Ave. Cos. dist. (to predicted signatures)	0.0793	0.0588	0.0462	0.0138
hyperparameter $\alpha_k$	0.01	0.1	1	10
# predicted signatures	10	10	10	6
Ave. Cos. dist. (to COSMIC Signatures)	0.0134	0.0138	0.0558	0.0480
Ave. Cos. dist. (to predicted signatures)	0.0134	0.0138	0.0588	0.1046

Note: In this simulation, we applied VB-LDA to a simulated mutation dataset in which 10 known COSMIC signatures are regarded as true signatures. The default setting of parameters is  $S=1000$ ,  $n_s=2000$  and  $\alpha_k=0.1$ , and we changed each parameter from its default value. ‘# predicted signatures’ (closer to 10 is better) is the number of predicted mutation signatures according to VB-LDA based on VLB (cf. Fig. 3); ‘Ave. Cos. dist. (to COSMIC Signatures)’ (smaller is better) is equal to  $d(\mathcal{P}, \mathcal{K}) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \arg \min_{k \in \mathcal{K}} \text{cos\_dist}(p, k)$  where  $\mathcal{P}$  (resp.  $\mathcal{K}$ ) is a set of predicted (resp. known) mutation signatures, and  $\text{cos\_dist}$  is cosine distance between two signatures, whereas ‘Ave. Cos. dist. (to predicted signatures)’ (smaller is better) is equal to  $d(\mathcal{K}, \mathcal{P}) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \arg \min_{p \in \mathcal{P}} \text{cos\_dist}(p, k)$ . When ‘# of predicted signatures’ is equal to 10 and the values of ‘Ave. Cos. dist. (to COSMIC signatures)’ and ‘Ave. Cos. dist. (to predicted signatures)’ are closer to 0, the correct signatures are assumed to be reproduced. The bold values indicate the best values within each row.



**Fig. 3.** Variational lower bound (VLB) for each number of mutation signatures in the simulation with an appropriate condition ( $S=1000$ ,  $n_s=2000$ , and  $\alpha_k=0.1$ ; cf. Table 1). In this analysis, the true number of signatures is 10. The horizontal axis shows the number of signatures  $K$ , and vertical axis indicates VLB for each  $K$ . Furthermore, red and green bars show the highest and the second highest VLBs, respectively. On the basis of VLB, we can select a true number of mutation signatures (i.e. 10)

estimated mutational signatures for  $K=10$ , respectively, with appropriate parameters ( $S=1000, n_s=2000, \alpha_k=0.1$ ). These results clearly mean that VB-LDA successfully estimated the correct number of mutation signatures ( $K=10$ ) and a mutational distribution for most conditions.

On the other hand, VB-LDA failed to estimate the true mutational distribution and the true number of mutation signatures when the number of samples was less than 100 and the number of mutations included in one sample ( $n_s$ ) was less than 400. This is presumably because VB-LDA estimates a signature on the basis of co-occurrence of mutations, which assumes a relatively large input dataset (Blei *et al.*, 2003). Besides, when hyperparameter  $\alpha$  is smaller than 0.1 (i.e. the distribution of mutational processes in the sample is sufficiently biased), our method successfully estimated the number of mutation signatures and the original distributions. We apply these observations to the selection of real data in the next section.

### 3.2 Results on real mutation datasets

To look for new mutation signatures, we predicted real mutation signatures by applying our method to the mutation catalog provided in the COSMIC database. Because a mutational process varies depending on the cancer type, samples were subdivided into primary lesions, which are used to train the LDA model.

Many mutation catalogs are registered in the COSMIC database (<https://cancer.sanger.ac.uk/cosmic/download>), but there are many samples that include only a small number of mutations (i.e.  $n_s$  is very small). We confirmed via the simulation that such samples are not appropriate for learning mutation signatures as discussed in Section 3.1; therefore, we filtered out samples with fewer than 400 mutations. In addition, the primary lesion whose number of suitable samples is less than 25 was excluded from the analysis for the same reason. As a result, 1607 samples in 11 categories of primary lesions (breast, endometrium, large intestine, liver, lung, oesophagus, prostate, skin, soft tissue, stomach, upper aerodigestive tract and urinary tract) were analyzed as follows. For some categories of primary lesions, the number of samples is small (e.g. 30 for breast cancer), and it is unclear whether our method can estimate the correct number of signatures from the simulation results. However, in this case, because the samples that have almost no mutations behave like noise, we decided to use only the sample with many mutations even if the number of samples decreases.

Then, based on four mutation dictionaries ( $\mathcal{M}_d, d=1, 2, 3, 4$ ), four datasets were constructed for each type of primary lesion. At this time, it is necessary to determine the surrounding bases at the site where the mutation occurred (mutation context) according to the corresponding mutation dictionary, and we used GRCh38 as a reference human genome to obtain it and produced mutation datasets. We provided the mutational burden of the samples used in the experiment and the information on site subtype and histology as Supplementary HTML materials (<http://www.f.waseda.jp/mhamada/MS/index.html>). Ramazzotti *et al.* (2018) pointed out that the mutational distributions are affected by the sequencing strategy (i.e. whole-genome or whole-exome sequencing). In this study, we did not care about the strategies because the mutational distributions were not different between the whole-sequence group (1575 samples) and targeted-sequencing group (32 samples) (the cosine distance between them is 0.1024; Supplementary Fig. S1). However, this might be because the majority of whole-sequence samples is whole-exome-sequenced samples. In that case, it is better to subdivide the samples having a whole-exome sequence and whole-genome sequence in the same way as in another study (Alexandrov *et al.*,

2013a), but the samples with whole-genome sequence are still too few to perform analyses; therefore, in this study we decided to carry out experiments without separating the samples.

Finally, we applied our LDA method to these mutation datasets and determined mutation signatures by VB inference (Section 2.3).

#### 3.2.1 VB-LDA successfully recovered known mutation signatures

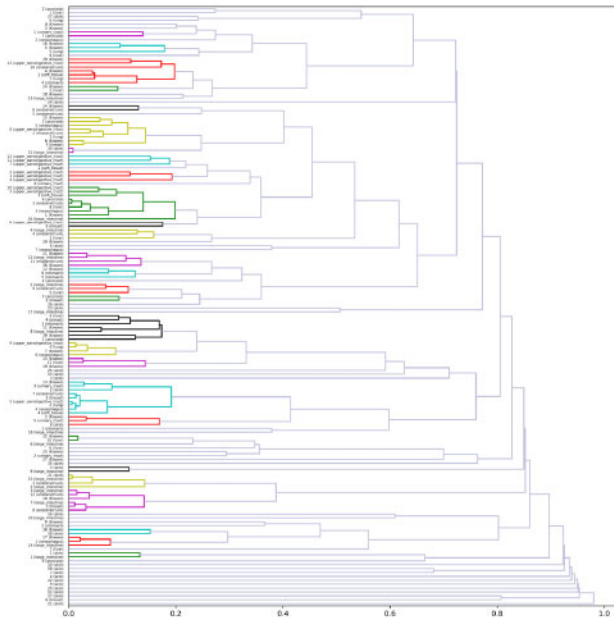
Supplementary Tables S1–S4 show a summary of the number of predicted signatures, which are similar to COSMIC Known Signatures for each mutation dictionary  $\mathcal{M}_k$  ( $k=1, 2, 3, 4$ ), suggesting that most of the known signatures in the COSMIC database were extracted by VB-LDA. Especially, when the mutation datasets with  $\mathcal{M}_3$  and  $\mathcal{M}_4$  (both of which include indels in their mutation vocabularies) were tested, our method successfully predicted mutation signatures that are likely to be associated with insertions and deletions (e.g. COSMIC Known Signatures 6 and 15). This result therefore supports the usefulness of various mutation dictionaries.

On the other hand, our method did not find several COSMIC Known Signatures with any dictionary. This is because we did not take into account all the available mutation catalogs in COSMIC. For example, our method did not extract COSMIC Known Signature 27, which has been found in kidney cancer (Alexandrov *et al.*, 2013b). Because there is a only small number of samples of kidney cancer that have more than 400 mutations in COSMIC, kidney cancer was removed from the list of targets of the experiment to perform stable modeling. Moreover, Supplementary Figure S3 indicates that our method estimated a merged mutation signature (predicted signature 1, a dominant signature in most samples; Supplementary Fig. S3A) of COSMIC signature 2 and 13 (Supplementary Fig. S3D). It is known that both COSMIC signatures 2 and 13 are related to AID/APOBEC proteins and often co-occur. From a methodological viewpoint, it would be difficult to deconvolute signatures that often co-occur (the convexity of transition of VLB is unclear; Supplementary Fig. S3B). Note that signatures in the COSMIC database are taken from several studies, and manually curated.

#### 3.2.2 Discovering new signatures on the basis of hierarchical clustering

To find a novel set of mutation signatures, we performed hierarchical clustering analyses for predicted mutation signatures with each mutation dictionary. In these analyses, we put predicted signatures all together regardless of the cancer types (i.e. primary lesion), and subdivided those signatures into clusters based on cosine distance, by the average-linkage hierarchical clustering method; the COSMIC Known Signatures were also included for dictionary  $\mathcal{M}_1$ . After clustering, we regarded a group of signatures as a cluster when cosine distances among members were less than 0.2 (Fig. 4 and Supplementary Figs S4–S7). It should be noted that mutation signatures that are found in several cancer types support the existence. Additionally, to ascertain the reliability of the signatures, signature activity,  $\theta_{s,k}$  in Section 2.2, is calculated, which shows how much relative contribution is present in a sample  $s$  for a specific signature  $k$ .

In the figures below, note that all bar graphs show a mutational distribution (i.e.  $\phi_k$ ). The horizontal axis means the mutation types, and the vertical axis shows their proportion of that signature as do Figure 1 and Supplementary Figure S1. Furthermore, red, green, blue, cyan, magenta and yellow bars show the probability of emergence of substitutions [C > A], [C > G], [C > T], [T > A], [T > C] and [T > G], respectively.

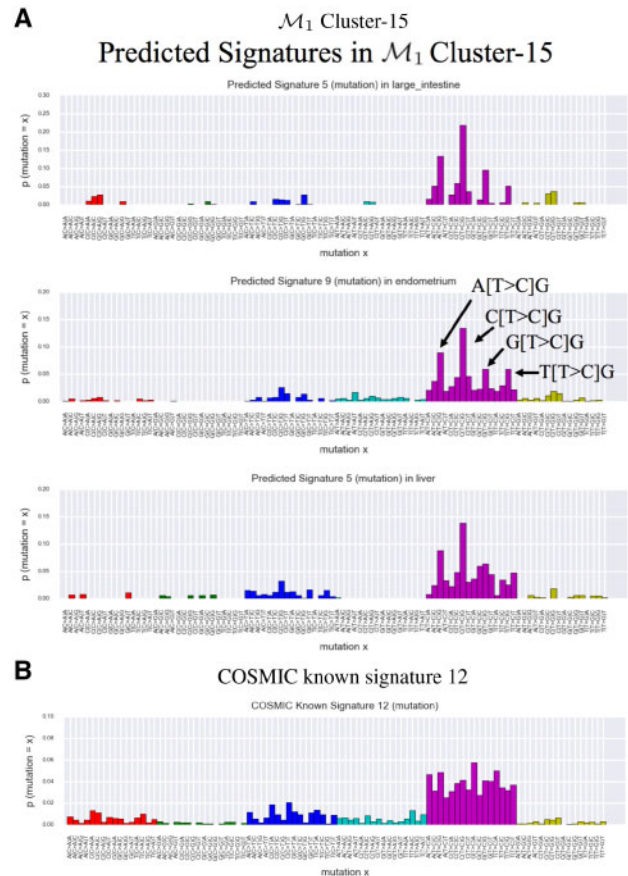


**Fig. 4.** Clustering results on mutation signatures for dictionary  $\mathcal{M}_1$ , where all the known COSMIC signatures in addition to our predicted signatures are included on the vertical axis. Averaged linkage hierarchical clustering with a cosine distance is employed for clustering. The colors show a cluster with the distance threshold 0.2. A high-quality image is available in [Supplementary Fig. S4](#) or Supplementary HTML material. The results for the other mutation dictionaries are presented in [Supplementary Figures S5–S7](#)

In the text below, we report seven interesting mutation signatures that might be novel, or otherwise broaden our understanding of known COSMIC signatures from the clustering results. Because they appear in multiple primary lesions, and most of the signatures are active in many samples (cf. [Supplementary Fig. S8](#)), the existence of these signatures is reliable. Higher-resolution figures for all the mutation signatures, clustering results and signature activity (i.e. the relative contribution of signatures in each sample) are shown in Supplementary HTML materials (<http://www.f.waseda.jp/mhamada/MS/index.html>). In the following, ‘ $\mathcal{M}_d$  Cluster-ID’ denotes the result of Cluster-ID with the  $\mathcal{M}_d$  dictionary (the ID of the cluster corresponds to the ID in Supplementary HTML).

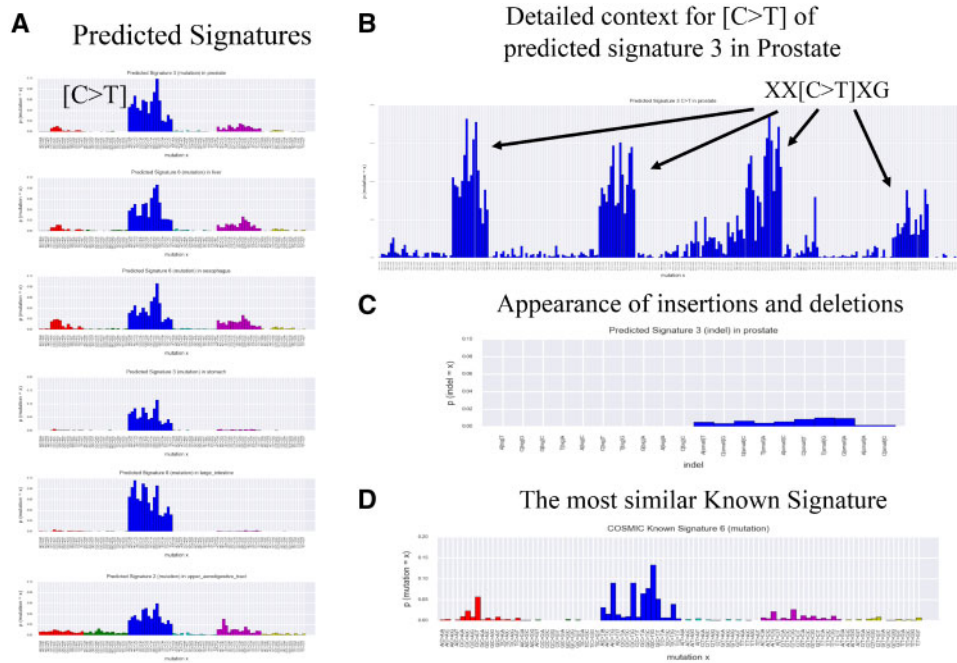
1.  $\mathcal{M}_1$  Cluster-15 ([Fig. 5A](#)): The signatures in this cluster were detected in the large intestine, endometrium and liver. We observed characteristic peaks,  $X[T>C]G$  ( $X = A, T, G$  or  $C$ ), in all the mutation signatures ([Fig. 5](#)), which might be influenced by an unknown mutational process. This cluster includes novel mutation signatures because there is no known signature that is similar to these signatures. For example, the most similar known signature for predicted signature 9 in the endometrium (the middle panel in [Fig. 5A](#)) is COSMIC Known Signature 12 (found in liver cancer), and the cosine distance between them is relatively large (0.2234). Note that ‘COSMIC known signature 12’ does not include the characteristic peaks,  $X[T>C]G$  ([Fig. 5B](#)).

On the other hand, this cluster could be related to COSMIC Known signature 1 for the following reason. The  $[TC]G$  peaks in this signature are a mirror image of the  $[CT]G$  peaks in COSMIC Known signature 1. If the reference genome (GRCh38) includes many  $[CT]$  substitutions affected by Known Signature 1, the  $[TC]$  substitutions have occurred in the other samples that were not affected by Known Signature 1. If this hypothesis is correct, then this signature does not represent a novel mutation process.



**Fig. 5.**  $\mathcal{M}_1$  Cluster-15. This cluster in panel (A) includes three mutation signatures from the large intestine (top), endometrium (middle) and liver (bottom) with the  $\mathcal{M}_1$  dictionary. The horizontal axis indicates the mutation types in  $\mathcal{M}_1$ , and the vertical axis shows their proportion estimated by our method (cf. [Fig. 1](#)). The arrows (in the middle panel) show four peaks in the mutation distribution at  $X[T>C]G$  ( $X = A, T, G$  or  $C$ ), which are also observed in the other figures. Panel B illustrates the known COSMIC signature that is the known signature closest to the signatures in  $\mathcal{M}_1$  Cluster-15

2.  $\mathcal{M}_2$  Cluster-1 ([Supplementary Fig. S9A](#)) &  $\mathcal{M}_4$  Cluster-12 ([Supplementary Fig. S9B](#)). Both clusters are composed of two signatures, which are found in the stomach and oesophagus. Both signatures have peaks at  $C[T>C]X$  and  $C[T>G]X$  ( $X = A, T, G$  or  $C$ ), and the second and fourth panels in [Figure 9B](#) suggest that they are not related to indels. The detailed mutation context (up to 2 bases from the mutated base) shows that there is a high proportion of  $XC[T>G]XT$  ([Supplementary Fig. 9C](#)). Of note, these signatures were extracted from adjacent organs such as the stomach and oesophagus.
3.  $\mathcal{M}_2$  Cluster-2 &  $\mathcal{M}_4$  Cluster-11 ([Supplementary Fig. 10A](#)).  $\mathcal{M}_4$  Cluster-11 includes signatures from many organs (breasts, endometrium, large intestine and stomach), and they have peaks at  $T[C>A]X$  and  $T[C>T]X$  in common ( $X = A, G, C$  or  $T$ ). Furthermore, a similar cluster was detected when dictionary  $\mathcal{M}_2$  was chosen ( $\mathcal{M}_2$  Cluster-2). If we study the mutation context in detail, it offers distinct peaks at  $TT[C>A]XT$  and  $TT[C>T]XG$  ([Supplementary Fig. 10B](#)). Although  $\mathcal{M}_4$  is a mutation dictionary including indels, signatures belonging to this cluster were not related to indels. In the large intestine, two similar signatures are predicted and belong to this cluster (Predicted Signatures 2 and 7; the 3rd and 4th panels in [Supplementary Fig. S10A](#)). One



**Fig. 6.**  $\mathcal{M}_4$  Cluster-3. (A) Six predicted mutation signatures (for the prostate, liver, oesophagus, stomach, large intestine and upper aerodigestive tract) with respect to the substitution in this cluster, where mutation context of substitutions in  $\mathcal{M}_4$  is transformed to that of  $\mathcal{M}_1$  as described in Section 2.4, and the horizontal axis denotes the type of mutations in  $\mathcal{M}_1$ . All the signatures have peaks at the [C > T] mutation (blue bars in A), and mutation [C > A] or [T > C] is seen in some signatures. (B) In the figure showing the context in detail, the horizontal axis means a mutation context of [C > T] substitutions in  $\mathcal{M}_2$ . In particular, when we see the mutation context for up to 2 bases from a mutation, large peaks were observed at XX[C > T]XG as the four arrows indicate. (C) In the indel graph, the horizontal axis means mutation types of  $\mathcal{M}_1$ , and signatures extracted from the oesophagus, liver, prostate and upper aerodigestive tract with  $\mathcal{M}_4$  have indels as their cause. (D) In addition, the bar graph presents COSMIC Known Signature 6, which is the most similar known signature (e.g. cosine distance between it and Predicted Signature 5 in the oesophagus with  $\mathcal{M}_2$  is 0.1296)

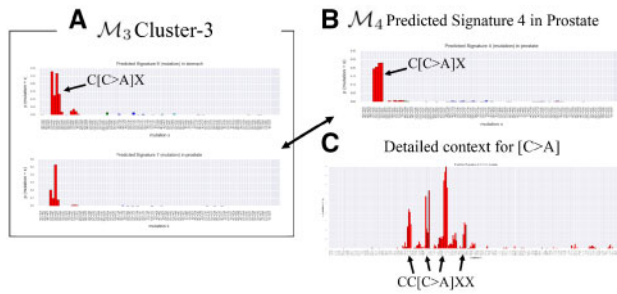
possibility is that these signatures were obtained via a model selection mistake, but the relative contributions of these two signatures are clearly different among samples (Supplementary Fig. S11), thus pointing to the existence of a mutational process with multiple kinds of mutational distributions such as AID/APOBEC family (i.e. Cosmic Known Signatures 2 and 13). Besides, from the appearance of peaks, these signatures are probably related to COSMIC Known Signature 10 associated with POLE defects. If we look at activities (Supplementary Fig. S8), there are many cases in which these signatures are dominant in large-intestine and endometrium samples. Variants of cancer-related genes, POLE and POLD1, have been found in patients with large-intestine or endometrial cancer (Briggs and Tomlinson, 2013; Palles et al. 2013); therefore, the relation between these signatures and COSMIC Known Signature 10 is obvious.

4.  $\mathcal{M}_2$  Cluster-6 &  $\mathcal{M}_4$  Cluster-3 (Fig. 6). The mutation signatures in  $\mathcal{M}_4$  Cluster-3 are found in many organs such as the prostate, liver, oesophagus, stomach, large intestine and the upper aerodigestive tract (Fig. 6A). When the  $\mathcal{M}_2$  dictionary is used, a signature is obtained that is similar to Cluster-6 with  $\mathcal{M}_2$  (Supplementary HTML materials). All the signatures have peaks at the [C > T] substitution in their mutation distribution, and mutation [C > A] or [T > C] is seen in some signatures. In particular, when we see the mutation context for up to 2 bases from a mutation, characteristic peaks were observed at XX[C > T]XG (Fig. 6B). Note that mutation signatures extracted from the oesophagus, liver, prostate and upper aerodigestive tract with  $\mathcal{M}_4$  have a preference for indels (Fig. 6C). Because they are related to indels and there are peaks at substitutions [C > T], we believe that these signatures have some relation to COSMIC Known

Signature 6, which is associated with defective DNA mismatch repair (Fig. 6D). Actually, the cosine distance between any members belonging to this cluster and COSMIC Known Signature 6 is sufficiently small (e.g. cosine distance between Predicted Signature 5 in the oesophagus with  $\mathcal{M}_2$  and COSMIC Known Signature 6 is 0.1296). If this signature group is related to defective DNA mismatch repair, then this cluster shows an extended sequence context preference for COSMIC Known Signature 6.

5.  $\mathcal{M}_2$  Cluster-7 (Supplementary Fig. S12) &  $\mathcal{M}_4$  Cluster-4.  $\mathcal{M}_2$  Cluster-7 (Supplementary Fig. S12) is associated with  $\mathcal{M}_4$  Cluster-4 (shown in the Supplementary HTML material). The signatures in  $\mathcal{M}_2$  Cluster-7 are found in the stomach, large intestine and endometrium (Supplementary Fig. S12). A similar signature was also extracted from lungs in  $\mathcal{M}_4$  Cluster-4. Every signature in  $\mathcal{M}_2$  Cluster-7 has peaks at G[C > T]X, and the signature of the endometrium has peaks at A[C > T]X (Supplementary Fig. S12B). Additionally, mutation signatures from the stomach and large intestine have definite peaks at GG[C > T]XG and TG[C > G]XG in particular (Supplementary Fig. S12B). Besides, any member of  $\mathcal{M}_4$  Cluster-4 corresponding to this cluster has no indels.

6.  $\mathcal{M}_2$  Cluster-9 (Supplementary Fig. S13) &  $\mathcal{M}_4$  Cluster-8 (Supplementary Fig. S14).  $\mathcal{M}_2$  Cluster-9 is associated with  $\mathcal{M}_4$  Cluster-8. In the results on  $\mathcal{M}_2$ , these signatures were found in the upper aerodigestive tract, lungs and skin (Supplementary Fig. S13A). All the estimated mutation signatures have peaks at C[C > T]X and T[C > T]X, and this tendency is particularly strong when the 3' adjacent base for the mutation is thymine (i.e. the peaks of C[C > T]C and T[C > T]C are stronger than the other peaks). If we examine the mutation context in detail, there



**Fig. 7.**  $\mathcal{M}_3$  Cluster-3. (A) Two mutation signatures (from the stomach and prostate) included in this cluster; the arrow points to peaks at C[C>A]A and C[C>A]G. (B) The (marginalized) Predicted Signature 4 in the prostate (with the  $\mathcal{M}_4$  dictionary) similar to the signatures in  $\mathcal{M}_3$  Cluster-3, where the arrow indicates the peaks at C[C>A]X (X = A, G, C or T). (C) The detailed mutation context (up to 2 bases from the mutation) for the C[C>A]X peaks in the mutation signature shown in (B); the arrows indicate CC[C>A]XX peaks (X = A, G, C and T)

are peaks at XT[C>T]XC (Supplementary Fig. S13B). Furthermore, this cluster may be meaningful because all primary lesions in which these signatures are found are in epithelial tissue and have similar peaks in comparison with COSMIC Known Signature 7 (Supplementary Fig. S13C), which is associated with exposure to ultraviolet light (e.g. cosine distance between Predicted Signature 2 from lungs with  $\mathcal{M}_2$  and COSMIC Known Signature 7 is 0.2076, which is relatively small). Therefore, the mutational process of these signatures may be related to exposure to ultraviolet light. In particular, Predicted Signature 2 in skin with  $\mathcal{M}_2$  has a stable relative contribution through samples (Supplementary Fig. S8).

- $\mathcal{M}_3$  Cluster-3 (Fig. 7). Two mutation signatures in  $\mathcal{M}_3$  Cluster-3 were found in the stomach and prostate, and they have peaks at C[C>A]X (X = A, C, G or T), especially at C[C>A]A and C[C>A]G (Fig. 7A). The signatures in this cluster are novel because, for example, the known signature most similar to Predicted Signature 7 for the prostate (the bottom panel in Fig. 7A) is COSMIC Known Signature 24, whose cosine distance is 0.7252, and they are clearly different signatures. We found that the signatures in the  $\mathcal{M}_3$  cluster are similar to Predicted Signature 4 with  $\mathcal{M}_4$  (Fig. 7B), which has strong peaks at CC[C>A]XX when we see the mutation context in detail (Fig. 7C), indicating that the 5' base is important for this mutation.

### 3.3 Further discussion

In this study, we proposed a method for estimating mutation signatures; the advantage of this method is the mathematical validity in model selection (i.e. selecting the number of mutation signatures), compared with existing methods that involve ad hoc approaches or improper approximation for model selection. To test whether this mathematical justification is effective for the estimation of the number of signatures and validity of the obtained signatures, we compared VB-LDA with the existing method EMu and probabilistic latent semantic analysis (PLSA), which is the probabilistic model that Dirichlet distribution of a prior is excluded from LDA. Details are given in the Supplementary material (Supplementary Section 4). In summary, from the viewpoint of model selection, we confirmed via the simulation that VB-LDA is superior to EMu (if a mutation opportunity is not considered) and to PLSA. Nevertheless, in

benchmarking using real data, it was not possible to compare the effectiveness because the true signature set could not be known.

In addition to the signatures listed in Section 3.2.2, novel mutation signatures that are consistently active in multiple samples were found. In the clustering method (cf. Section 3.2.2), more reliable signatures are selected on the basis of the criteria that similar signatures are obtained from multiple independent primary lesions. Nevertheless, there may exist mutational processes that are active only in a specific primary lesion, and the reliability of those signatures can be indicated by high activity in multiple samples. Specifically, Signature 6 in the lung with  $\mathcal{M}_1$  and Signature 1 in the endometrium with  $\mathcal{M}_2$  are mutation signatures whose median and average values of activities were greater than 0.05 and whose cosine distance to known signatures was greater than 0.2 (Supplementary Fig. S15). The former has peaks at C[C>A]A and T[C>A]G, and the latter has peaks at XX[C>A]XT (Supplementary Fig. S15). Especially signature 1 in the endometrium with  $\mathcal{M}_2$  (Supplementary Fig. S15C) forms a cluster with signature 5 in the large intestine with  $\mathcal{M}_2$ , and the activity also shows a moderate value (the median and means are 0.0563 and 0.0413).

It is widely known that different signatures are obtained for different tissues even with the same primary lesion (Alexandrov et al., 2013b). In the present study, we did not conduct analysis by further distinguishing the mutation catalog of the same primary lesion by histological or a site subtype information because the number of samples included in one dataset becomes insufficient to apply our method (cf. Section 3.1). For example, investigating the signature activity of soft tissue with  $\mathcal{M}_2$  dictionary (Supplementary Fig. S16), four signatures were predicted and only one signature is dominantly active in each sample that corresponds to one of the three site subtypes (fibrous tissue of uncertain origin, striated muscle and blood vessels). Therefore, to make effective use of those information, increasing the number of samples and using the resampling technology typified by the bootstrap method may be essential to avoid overfitting.

This study uncovered multiple signatures that cause indels as well as substitutions (e.g.  $\mathcal{M}_4$  Cluster-3). Nonetheless, it has been shown that indel-generating processes depend on not only sequence context/indel length but also the features of the surrounding sequence such as overlapping micro-homology (Alexandrov et al., 2018; Nik-Zainal et al., 2016), which cannot be captured by conventional models. In addition, we did not include variants other than substitutions and indels (e.g. rearrangement) in the vocabulary of a mutation dictionary. Particularly, in recent years, there have been reports of signatures in which a rearrangement is present specifically (Nik-Zainal et al., 2016). Thus, to elucidate the whole mutation signature, it is necessary to pay attention to these mutations as well as substitutions and indels.

Furthermore, it may be necessary to review how to analyze the mutation context. In our study (and other existing studies), only two bases upstream and downstream of the mutated base were incorporated into the mutation context. With  $\mathcal{M}_2$  or  $\mathcal{M}_4$ , there were many mutation signatures depending on the mutation context upstream or downstream by 2 bases from the mutated base [e.g. Predicted Signature 3 in the prostate with  $\mathcal{M}_4$  (Fig. 6B)], suggesting that a signature depends on the mutation context 3 bases or more away from the mutated base. From the viewpoint of sparseness of data, it is difficult to analyze samples with a new mutation dictionary taking into account longer mutation contexts as compared to  $\mathcal{M}_2$  and  $\mathcal{M}_4$ . For example, if 3 bases upstream and downstream of a mutated base are included in the mutation context, the size of that mutation



dictionary is equal to 24 576 for substitutions. Further studies should be conducted regarding how to define more complicated and informative mutation dictionaries (e.g. by considering longer mutation contexts).

Our computational study indicates that our method not only successfully estimates the number of mutation signatures on simulated data but also predicts several new mutation signatures whose reliability is provided by signature activity and clustering of signatures. On the other hand, the computational experiments also clarify several limitations of the proposed method (e.g. some known signatures could not be recovered; more than two very similar signatures are estimated in one cancer type), and further studies are necessary for improvement of the method. In particular, to determine one mutation signature corresponding to one mutational process, it is necessary to select the ‘representative’ signature or merge multiple ones obtained from all the primary lesions in which that mutational process acts. Although the mutational distribution of the extracted signature may be slightly different, we should not separate samples for each primary lesion but rather analyze the samples collectively to solve this problem. Nevertheless, in this case, the complexity of the dataset increases, making identification of the signatures difficult. Therefore, we are now devising a new Bayesian hierarchical model to extract mutation signatures where the hyperparameters of a prior distribution in LDA capture the features of each primary lesion (Supplementary Fig. S20). By means of such a model, it becomes possible to analyze samples collectively, with sharing of mutational distributions of signatures among primary lesions. We expect that this model will produce a set of signatures that do not include multiple signatures from the same mutational process.

## 4 Conclusion

In this study, we proposed an effective method for determining the number and characteristics of mutation signatures by VB-LDA, and discovered many interesting mutation signatures, which might be novel, for mutation data of various cancer genomes with four mutation dictionaries. Compared with other existing approaches, our method employs a different approach for model selection, thereby leading to different mutation signatures. Furthermore, we introduced a hierarchical clustering procedure for predicted mutation signatures in addition to known signatures and obtained better information about which mutation signature appears in which primary lesion and where peaks appear in specific mutation signatures. All the predictions are freely available on our website (<http://www.f.waseda.jp/mhamada/MS/index.html>), which will be a useful resource for cancer research.

Although we discovered interesting mutation signatures in this study, we could not determine biological implications of the newly found mutation signatures (note that mutational processes for most of known signatures in the COSMIC database are unknown). In the future, we should consider the correspondence between mutation contexts of these signatures and actual etiology. These data will lead to further understanding of carcinogenic mechanisms and early diagnosis of cancers.

## Acknowledgements

Computation for this study was partially performed on the NIG supercomputer at ROIS National Institute of Genetics. YU is currently working at a pharmaceutical company in Japan.

## Funding

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) [KAKENHI grant numbers JP18KT0016, JP17K20032, JP16H05879, JP16H01318 and JP16H02484 to M.H.] JST CREST Grant Number JPMJCR1881, Japan and by a Waseda University Grant for Special Research Projects (Project Number: 2017A-506).

*Conflict of Interest:* none declared.

## References

- Alexandrov, L. *et al.* (2018) The repertoire of mutational signatures in human cancer. *bioRxiv*, 322859.
- Alexandrov, L.B. *et al.* (2013a) Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.*, **3**, 246–259.
- Alexandrov, L.B. *et al.* (2013b) Signatures of mutational processes in human cancer. *Nature*, **500**, 415.
- Alexandrov, L.B. *et al.* (2015) Clock-like mutational processes in human somatic cells. *Nat. Genet.*, **47**, 1402.
- Barba, M. *et al.* (2014) Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses*, **6**, 106–136.
- Blei, D.M. *et al.* (2003) Latent dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993–1022.
- Briggs, S. and Tomlinson, I. (2013) Germline and somatic polymerase  $\epsilon$  and  $\delta$  mutations define a new class of hypermutated colorectal and endometrial cancers. *J. Pathol.*, **230**, 148–153.
- Corduneanu, A. and Bishop, C.M. (2001) Variational Bayesian Model Selection for mixture distributions. *Artificial Intelligence and Statistics 2001*, 27–34.
- Fischer, A. *et al.* (2013) Emu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol.*, **14**, R39.
- Flaherty, P. *et al.* (2005) A latent variable model for chemogenomic profiling. *Bioinformatics*, **21**, 3286–3293.
- Forbes, S.A. *et al.* (2015) Cosmic: exploring the world’s knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.
- Greenman, C. *et al.* (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, **446**, 153–158.
- Harris, R.S. (2013) Cancer mutation signatures, dna damage mechanisms, and potential clinical implications. *Genome Med.*, **5**, 87.
- Hofmann, T. (1999) Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 50–57.
- Lee, D.D. and Seung, H.S. (2001) Algorithms for non-negative matrix factorization. In: *Advances in Neural Information Processing Systems*, pp. 556–562.
- Liu, B. *et al.* (2010) Identifying functional mirna–mrna regulatory modules with correspondence latent dirichlet allocation. *Bioinformatics*, **26**, 3105–3111.
- Nik-Zainal, S. *et al.* (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell*, **149**, 979–993.
- Nik-Zainal, S. *et al.* (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, **534**, 47.
- Palles, C. *et al.* (2013) Germline mutations affecting the proofreading domains of pole and pold1 predispose to colorectal adenomas and carcinomas. *Nat. Genet.*, **45**, 136.
- Pfeifer, G. (2006) *DNA Methylation: Basic Mechanisms*. Springer, Berlin, Heidelberg, pp. 259–281.
- Ramazzotti, D. *et al.* (2018) De novo mutational signature discovery in tumor genomes using sparsesignatures. *bioRxiv*, 384834.
- Rosales, R.A. *et al.* (2017) Signer: an empirical bayesian approach to mutational signature discovery. *Bioinformatics*, **33**, 8–16.
- Rubin, A.F. and Green, P. (2009) Mutation patterns in cancer genomes. *Proc. Natl. Acad. Sci. USA*, **106**, 21766–21770.
- Shiraishi, Y. *et al.* (2015) A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLoS Genet.*, **11**, e1005657.

- Smalheiser,N.R. (2002) Informatics and hypothesis-driven research. *EMBO Rep.*, **3**, 702.
- Stratton,M.R. (2011) Exploring the genomes of cancer cells: progress and promise. *Science*, **331**, 1553–1558.
- Stratton,M.R. et al. (2009) The cancer genome. *Nature*, **458**, 719–724.
- Temko,D. et al. (2018) The effects of mutational processes and selection on driver mutations across cancer types. *Nat. Commun.*, **9**, 1857.
- Tomczak,K. et al. (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. (Pozn)*, **19**, 68–77.
- Toyooka,S. et al. (2003) The tp53 gene, tobacco exposure, and lung cancer. *Hum. Mutat.*, **21**, 229–239.
- Wagner,R. et al. (2015) Analysis of mutational signatures in exomes from B-cell lymphoma cell lines suggest APOBEC3 family members to be involved in the pathogenesis of primary effusion lymphoma. *Leukemia*, **29**, 1612–1615.
- Wong,W.C. et al. (2011) Chasm and snvbox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics*, **27**, 2147–2148.
- Yamazaki,K. and Watanabe,S. (2005) Algebraic geometry and stochastic complexity of hidden markov models. *Neurocomputing*, **69**, 62–84.
- Zou,X. et al. (2018) Validating the concept of mutational signatures with isogenic cell models. *Nat. Commun.*, **9**, 1744.