Supplementary Information for “Discovering novel mutation signatures by latent Dirichlet allocation with variational Bayes inference”

Taro Matsutani, Yuki Ueno, Tsukasa Fukunaga, and Michiaki Hamada[*]

Contents

1 Learning parameters in LDA to model mutation signatures

During modeling of mutation signatures for LDA, the goal is to express observed mutation m well by inferring mutation signatures, in other words, the pair of latent variables and parameters $\{\boldsymbol{z}, \boldsymbol{\phi}\}$. It is desirable that these posterior problems are analytically solved, although we cannot solve the posterior distribution $p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{m}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ in practice. Consequently, under the assumption that factorization is possible, we aimed to construct an approximate distribution:

$$q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) := \prod_{s=1}^{S} \prod_{i=1}^{n_s} q(z_{s,i}) \prod_{s=1}^{S} q(\theta_s) \prod_{k=1}^{K} q(\phi_k) \tag{1}$$

KL divergence between an approximate distribution and true distribution is considered quality of approximate posterior distribution, but it includes the posterior distribution $p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{m}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ explicitly so that we cannot solve it. Thus, using the property for log-marginal likelihood about z:

$$\begin{aligned} &\mathbb{KL}[q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) \mid\mid p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{m}, \boldsymbol{\alpha}, \boldsymbol{\beta})] \\ &\quad = \log p(\boldsymbol{m} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) - F[q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi})] \end{aligned} \tag{2}$$

then, we can formulate an optimization problem for an approximate posterior distribution as follows.

$$q^*(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) := \underset{q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) \in \mathcal{Q}}{\arg\max}\ F[q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi})] \tag{3}$$

In the above formula, $\mathcal{Q}$ means the parameter set that can be factorized, and $F[q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi})]$ is called “variational lower bound: VLB” which means the lower bound of log-marginal likelihood $\log p(\boldsymbol{m} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$. KL divergence takes a non-negative value, and log-marginal likelihood is independent of the approximate posterior distribution, so that maximization of VLB is equal to minimization of KL divergence.

[*] To whom correspondence should be addressed. Tel: +81 3 5286 3130; Fax: +81 3 5286 3130; Email: mhamada@waseda.jp

According to Bayes' theorem, VLB is computed as follows:

$$
\begin{aligned}
&F[q(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})] \\
&= \int \sum_{\boldsymbol{z}} q(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}) \log \frac{p(\boldsymbol{m},\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi} \mid \boldsymbol{\alpha},\boldsymbol{\beta})}{q(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})} d\boldsymbol{\theta} d\boldsymbol{\phi} \\
&= \int \sum_{s=1}^{S} \sum_{i=1}^{n_s} q(z_{s,i}) q(\boldsymbol{\theta}_s) q(\boldsymbol{\phi}_k) \log p(m_{s,i} \mid z_{s,i},\boldsymbol{\phi}) p(z_{s,i} \mid \boldsymbol{\theta}_s) d\boldsymbol{\theta} d\boldsymbol{\phi} \\
&- \sum_{s=1}^{S} \sum_{i=1}^{n_s} \sum_{k=1}^{K} q(z_{s,i}=k) \log q(z_{s,i}=k) \\
&- \sum_{s=1}^{S} \mathbb{KL}[q(\boldsymbol{\theta}_s) \mid\mid p(\boldsymbol{\theta}_s \mid \boldsymbol{\alpha})] - \sum_{k=1}^{K} \mathbb{KL}[q(\boldsymbol{\phi}_k) \mid\mid p(\boldsymbol{\phi}_k \mid \boldsymbol{\beta})]
\end{aligned} \tag{4}
$$

where $q(\cdot)$ means participation of the functional argument. The approximate posterior distribution can be factorized, then we have to extract only items that are related to $q(\boldsymbol{z})$, $q(\boldsymbol{\theta})$ and $q(\boldsymbol{\phi})$ by means of equation (4) and implement the update parameter for maximization of VLB according to the variational method. Then we get the update formula:

$$
\begin{aligned}
&q(z_{s,i}=k) \\
&\propto \frac{\exp\left[\Psi(\xi^{\phi}_{k,m_{s,i}})\right]}{\exp\left[\Psi(\sum_{v'=1}^{V} \xi^{\phi}_{k,v'})\right]} \frac{\exp\left[\Psi(\xi^{\theta}_{s,k})\right]}{\exp\left[\Psi(\sum_{k'=1}^{K} \xi^{\theta}_{s,k'})\right]}
\end{aligned} \tag{5}
$$

$$
q(\boldsymbol{\theta}_s \mid \boldsymbol{\xi}^{\theta}_s) = \frac{\Gamma(\sum_{k=1}^{K} \xi^{\theta}_{s,k})}{\prod_{k=1}^{K} \Gamma(\xi^{\theta}_{s,k})} \prod_{k=1}^{K} \theta_{s,k}^{\xi^{\theta}_{s,k}-1} \tag{6}
$$

$$
q(\boldsymbol{\phi}_k \mid \boldsymbol{\xi}^{\phi}_k) = \frac{\Gamma(\sum_{v=1}^{V} \xi^{\phi}_{k,v})}{\prod_{v=1}^{V} \Gamma(\xi^{\phi}_{k,v})} \prod_{v=1}^{V} \theta_{k,v}^{\xi^{\phi}_{k,v}-1} \tag{7}
$$

where $\Psi(\cdot)$ means the Digamma function and $\Gamma(\cdot)$ means the Gamma function, respectively. Then, $\boldsymbol{\xi}^{\theta}$ and $\boldsymbol{\xi}^{\phi}$ are parameters when $q(\boldsymbol{\theta})$ and $q(\boldsymbol{\phi})$ are regarded as a Dirichlet distribution, they are calculated as follows.

$$
\begin{aligned}
\xi^{\theta}_{s,k} &= E_{q(\boldsymbol{z}_s)}[n_{s,k}] + \alpha_k \\
&= \sum_{i=1}^{n_s} q(z_{s,i}=k) + \alpha_k
\end{aligned} \tag{8}
$$

$$
\begin{aligned}
\xi^{\phi}_{k,v} &= E_{q(\boldsymbol{z})}[n_{k,v}] + \beta_v \\
&= \sum_{s=1}^{S} \sum_{i=1}^{n_s} q(z_{s,i}=k) \delta(m_{s,i}=v) + \beta_v
\end{aligned} \tag{9}
$$

2 Learning hyperparameters

In the above discussion, all priors are treated as already known, but actually it is desirable that hyperparameters can be learned by observation. In the same way as other parameters, it is natural to use VLB as the evaluation function for hyperparameter learning. Hence, we chose "fixed point iteration" for our method.

In fixed-point iteration, we search for fixed point x which fulfills $x = f(x)$ against function $f(x)$ from initial value x_0 by iterations. In practice, we should choose a lower bound for VLB and find the fixed point that maximizes it. The details are omitted, as we bring out items with regard to $\boldsymbol{\alpha}$, we obtain the lower bound of VLB:

$$
F[q(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})] \geq \sum_{s=1}^{S} \left[-b_s \sum_{k=1}^{K} \alpha_k + \sum_{k=1}^{K} a_{s,k} \log \alpha_k \right] + (\text{const.}) \tag{10}
$$

- $a_{s,k} = \left(\Psi(E[n_{s,k}] + \hat{\alpha}_k) - \Psi(\hat{\alpha}_k)\right)\hat{\alpha}_k$
- $b_s = \Psi\left(n_s + \sum_{k=1}^{K} \hat{\alpha}_k\right) - \Psi\left(\sum_{k=1}^{K} \hat{\alpha}_k\right)$

where $\hat{\alpha}$ means α before an update. Then, here is the update of $\boldsymbol{\alpha}$ to maximize the lower bound of VLB.

$$\alpha_k = \frac{\sum_{s=1}^{S}[\Psi(E[n_{s,k} + \hat{\alpha}_k]) - \Psi(\hat{\alpha}_k)]\hat{\alpha}_k}{\sum_{s=1}^{S}\left[\Psi(n_s + \sum_{k=1}^{K} \hat{\alpha}_k) - \Psi(\sum_{k=1}^{K} \hat{\alpha}_k)\right]} \tag{11}$$

Likewise, $\boldsymbol{\beta}$ is updated as

$$\beta_v = \frac{\sum_{k=1}^{K}[\Psi(E[n_{k,v} + \hat{\beta}_v]) - \Psi(\hat{\beta}_v)]\hat{\beta}_v}{\sum_{k=1}^{K}\left[\Psi(\sum_{v=1}^{V} E[n_{k,v}] + \hat{\beta}_v) - \Psi(\sum_{v=1}^{V} \hat{\beta}_v)\right]} \tag{12}$$

where $\hat{\beta}$ means β before an update as well as $\hat{\alpha}$. In summary, the algorithm that we used is shown as Algorithm 1.

Algorithm 1 Variational Bayes for LDA

```
Initialize randomly, $q(\boldsymbol{z})$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$   ($q(\cdot)$ means responsibility of a functional argument)
while  do
    Update $q(z_{s,i})$ $(1 \leq s \leq S,\ 1 \leq i \leq n_s)$ ◃ Eqs (5), (8) and (9)
    Update $q(\theta_s)$ $(1 \leq s \leq S)$ ◃ Eqs (6) and (8)
    Update $q(\phi_k)$ $(1 \leq k \leq K)$ ◃ Eqs (7) and (9)
    Update $\alpha_k, \beta_v$ $(1 \leq k \leq K,\ 1 \leq v \leq V)$ ◃ Eqs (11) and (12)
end while
```

3 Supplementary results

Note that higher resolution figures shown in this supplementary results are available at `http://www.f.waseda.jp/mhamada/MS/index.html`.

3.1 Results on simulated data

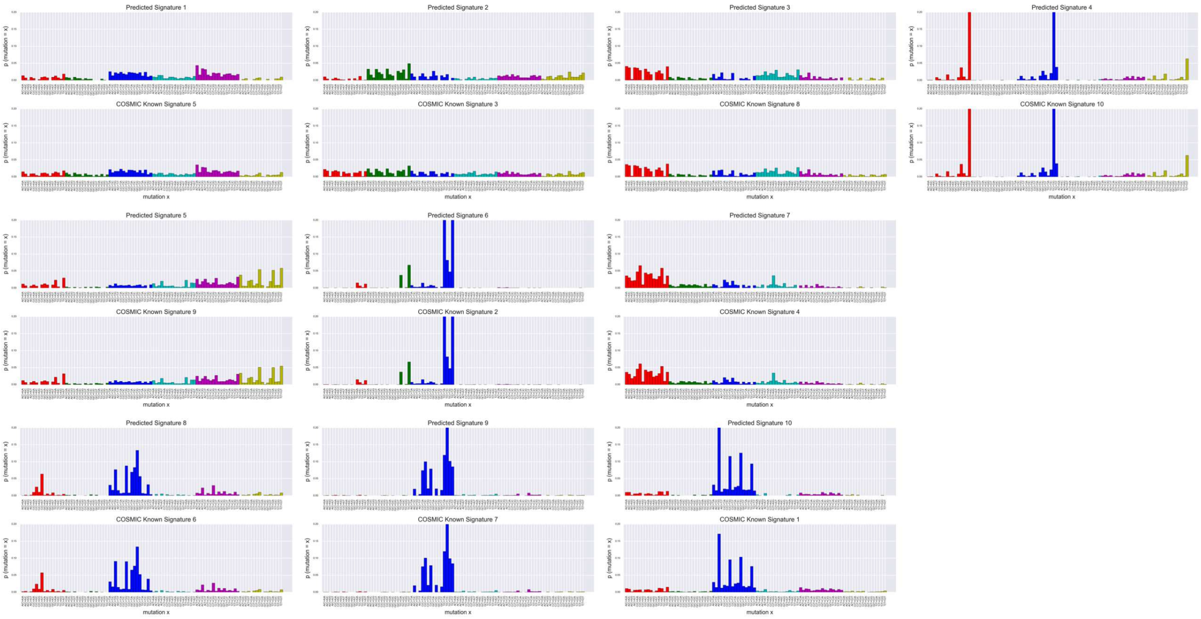


Figure S1: **All the predicted signatures and the corresponding COSMIC Known Signatures for simulated data with an appropriate condition ($S = 1000, n_s = 2000, \alpha_k = 0.1$).** The 1st, 3rd, and 5th rows indicate the mutation distributions of predicted mutation signatures by VB-LDA, whereas the 2nd, 4th, and 6th rows show the corresponding distributions of the known COSMIC signatures. Red, green, blue, cyan, magenta, and yellow bars show a set of [C>A], [C>G], [C>T], [T>A], [T>C], and [T>G] substitutions (in $\mathcal{M}_1$), respectively.

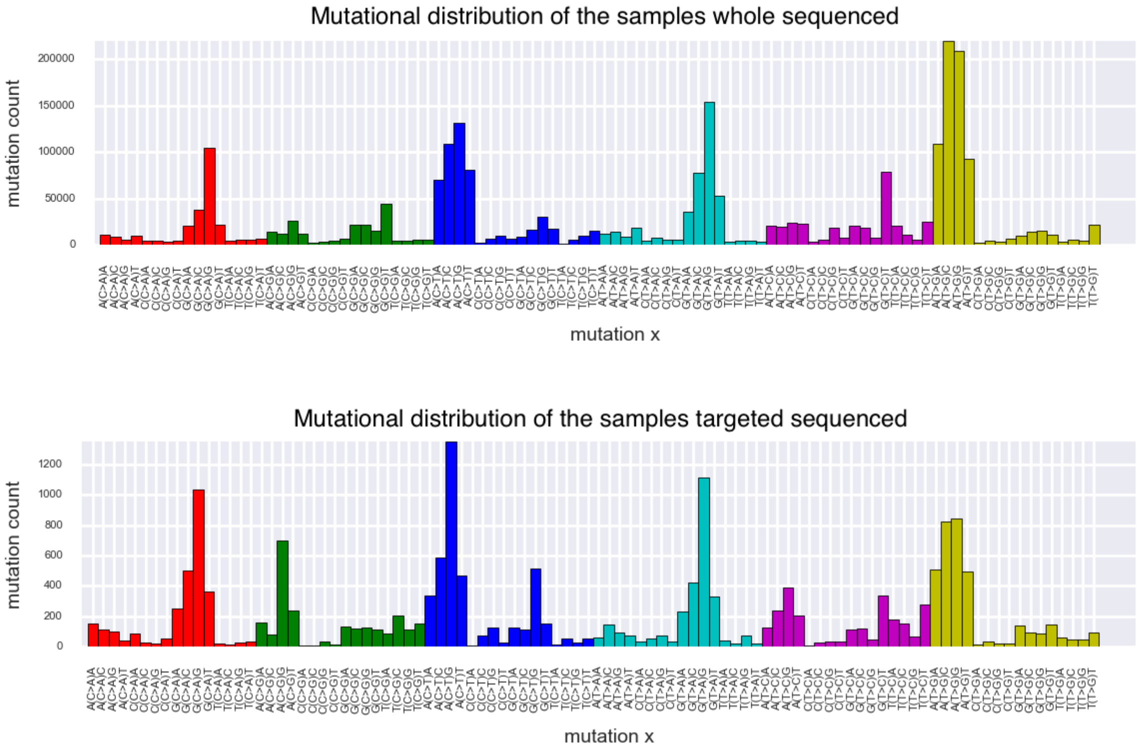


Figure S2: **Mutational distribution of the samples whole/targeted sequenced.** The upper graph shows the distribution of whole sequenced samples, and lower one shows that of targeted sequenced samples, respectively. In each graph, the horizontal axis shows the type of mutation, and the vertical axis shows the burden of that mutation. The cosine distance between whole/targeted sequenced samples was 0.1024.

3.2 A comparison of estimated mutation signatures with known COSMIC signatures

Table S1: The number of COSMIC Known Signatures that are similar to a predicted signature with dictionary $\mathcal{M}_1$.

Known	Breast	Endometrium	Large_intestine	Liver	Lung	Oesophagus	Prostate	Skin	Soft_tissue	Stomach	Upper_aerodigestive_tract	Urinary_tract	Total
1	0/0/1	1/1/1	1/1/1	1/1/1		1/1/1	1/1/1		1/1/1		3/3/3		9/9/10
2		1/1/1						1/1/1				1/1/1	3/3/3
3						0/1/1							0/1/1
4					1/1/1		1/1/1		1/1/1	1/1/1		1/1/1	5/5/5
5				0/1/1	1/1/1						1/1/3		2/3/5
6	1/1/1				1/1/1		0/1/1		1/1/1		0/1/2		3/5/6
7					1/1/1	1/1/1		0/1/1			1/1/1		3/4/4
8													
9													
10	1/1/1	2/2/2	2/2/2										5/5/5
11	1/1/1		1/1/1										2/2/2
12		0/1/1		0/1/1						2/2/2			2/4/4
13	1/1/1				1/1/1	1/1/1		1/1/1	0/1/1		1/1/1	1/1/1	6/7/7
14		1/2/2	0/1/1										1/3/3
15		1/1/1	0/1/1	1/1/1		1/1/1	1/1/1				1/1/1		5/6/6
16													
17			1/1/1			1/1/1		1/1/1					3/3/3
18													
19				1/1/1						1/1/1			2/2/2
20		0/1/1											0/1/1
21		1/1/1	1/1/1										2/2/2
22			0/1/1	1/1/1									1/2/2
23				1/1/1									1/1/1
24				1/1/1							1/1/1		2/2/2
25												0/0/1	0/0/1
26	1/1/1						1/1/1	0/1/1					2/3/3
27													
28													
29		1/1/1	1/1/1										2/2/2
30							1/1/1					0/0/1	1/1/2

Each cell $[i, j]$ represents the number of predicted signatures whose cosine distance toward COSMIC Known Signature i is the smallest among all COSMIC Known Signatures in cancer type j. The number $x/y/z$ means the number of predicted signatures whose cosine distances to Known Signatures are less than 0.2/0.25/0.3, respectively. Empty cells mean 0/0/0.

Table S2: The number of COSMIC Known Signatures that are similar to predicted signatures with the $\mathcal{M}_2$ dictionary.

Known	Breast	Endometrium	Large_intestine	Liver	Lung	Esophagus	Prostate	Skin	Soft_tissue	Stomach	Upper_aerodigestive_tract	Urinary_tract	Total
1			0/1/1										0/1/1
2	0/0/1		1/1/1					0/1/1					1/2/3
3						0/0/1							0/0/1
4				0/0/2	1/1/2			0/0/1	1/1/1				2/2/6
5	1/1/1		1/1/1	1/3/3				1/1/1	1/3/3	0/1/1	1/1/2		6/11/12
6		1/1/1		1/1/1	1/1/1	1/1/1	1/1/1			1/1/1	1/1/1		7/7/7
7					0/1/1			1/2/3			0/1/1	0/0/1	1/4/6
8													
9			0/1/1										0/1/1
10		0/1/1	1/1/1										1/2/2
11													
12		1/1/1		0/1/1									1/2/2
13													
14		0/0/1											0/0/1
15	1/1/1	1/1/1	2/2/2				1/1/1			1/1/1			6/6/6
16													
17													
18													
19	1/1/1		1/1/1	1/1/1									3/3/3
20										0/0/1			0/0/1
21			1/1/1							1/1/1			2/2/2
22				1/1/1									1/1/1
23													
24				1/1/1									1/1/1
25			0/0/1										0/0/1
26													
27													
28													
29													
30		1/1/1	0/0/1			0/1/1	0/1/1	1/1/1					2/4/5

The interpretation is the same as in Supplementary Table S1. How to associate predicted signatures with $\mathcal{M}_2$ (which have the information about surrounding bases of a mutated base for 2 bases ahead) to COSMIC Known Signatures is described in Section 2.4.

Table S3: The number of COSMIC Known Signatures that are similar to predicted signatures with the $\mathcal{M}_3$ dictionary.

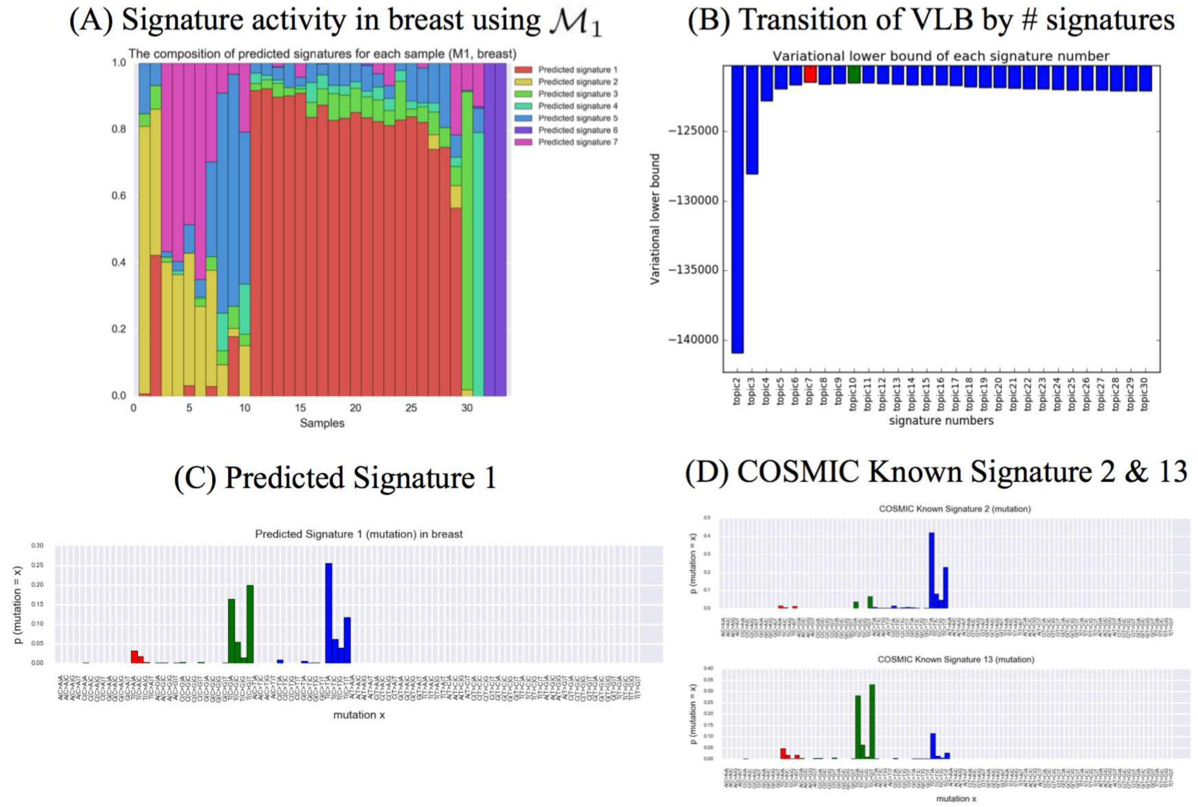
Known	Breast	Endometrium	Large_intestine	Liver	Lung	Oesophagus	Prostate	Skin	Soft_tissue	Stomach	Upper_aerodigestive_tract	Urinary_tract	Total
1	1/1/1	1/1/1	1/1/1	1/2/2		1/1/1	1/1/1	0/0/1	1/1/1	1/1/1	3/3/3		11/12/13
2	1/1/1	1/1/1						1/1/2				1/1/2	4/4/6
3			0/1/1				1/1/1			1/1/1			2/3/3
4					1/1/1				1/1/1				2/2/2
5					1/1/1						1/3/4	1/1/1	3/5/6
6	1/1/1	1/1/2	0/0/1	0/0/1	1/1/1		1/2/2		2/2/2	2/2/2			8/9/12
7					1/1/1			0/1/1			1/1/1		2/3/3
8													
9													
10	1/1/1	3/3/3	1/1/2							1/1/1			6/6/7
11			1/1/1										1/1/1
12		1/1/1						1/1/1					2/2/2
13	1/1/1				1/1/1	1/1/1		1/1/1	0/0/1		1/1/1	1/1/1	6/6/7
14		1/1/1	0/0/1										1/1/2
15		1/2/2	1/1/2	1/1/1		1/1/1				2/2/2	1/1/1		7/8/9
16				0/1/1									0/1/1
17						1/1/1		1/1/1		1/1/1			3/3/3
18			1/1/1										1/1/1
19	1/1/1												1/1/1
20			1/1/1										1/1/1
21			1/1/1							1/1/1			2/2/2
22			1/1/1	1/1/1									2/2/2
23				2/2/2									2/2/2
24		1/1/1		1/1/1									2/2/2
25												0/0/1	0/0/1
26	1/1/1												1/1/1
27													
28										1/1/1			1/1/1
29											1/1/1		1/1/1
30						0/1/1	1/1/1			1/1/1			2/3/3

The interpretation is the same as in Supplementary Table S1. How to associate the predicted signatures with $\mathcal{M}_3$ (which have not only substitutions but also indels as mutation types) to COSMIC Known Signatures is described in Section 2.4.

Table S4: The number of COSMIC Known Signatures that are similar to predicted signatures with the $\mathcal{M}_4$ dictionary.

Known	Breast	Endometrium	Large_intestine	Liver	Lung	Esophagus	Prostate	Skin	Soft_tissue	Stomach	Upper_aerodigestive_tract	Urinary_tract	Total
1			0/1/1										0/1/1
2	0/0/1		0/1/1					0/1/1					0/2/3
3										0/0/1			0/0/1
4					1/1/3			0/0/1	1/1/1				2/2/5
5		0/1/1	0/1/1	1/4/4					0/2/2	0/1/1	1/1/2	0/0/1	2/10/12
6		1/2/2	1/1/1	1/1/1	1/1/1	1/1/1	1/1/1	1/1/2		3/3/3			10/11/12
7					0/1/1			0/2/3			0/1/1	0/0/1	0/4/6
8													
9													
10		0/1/1	0/1/1										0/2/2
11													
12		1/1/1		1/1/1				1/1/1					3/3/3
13								0/0/1					0/0/1
14													
15	1/1/1	1/1/1	2/2/2				0/1/1		0/1/1	1/1/1			5/7/7
16													
17													
18													
19	1/1/1		1/1/1	1/1/1		0/1/1		0/0/1			1/1/1		4/5/6
20													
21			1/1/1							1/1/1			2/2/2
22				1/1/1									1/1/1
23													
24				1/1/1									1/1/1
25			0/1/1										0/1/1
26													
27													
28													
29													
30		0/1/1					0/1/1	0/1/1		1/1/1			1/4/4

The interpretation is the same as in Supplementary Tables S1, S2, and S3. How to associate the predicted signatures with $\mathcal{M}_4$ to COSMIC Known Signatures is described in Section 2.4.



(A) Signature activity in breast using $\mathcal{M}_1$

(B) Transition of VLB by # signatures

(C) Predicted Signature 1

(D) COSMIC Known Signature 2 & 13

Figure S3: (**A**) Signature activity in breast cancer with the $\mathcal{M}_1$ dictionary. The horizontal axis shows samples, and the vertical axis shows the activity of signatures in sample s $(= \boldsymbol{\theta}_s)$, where the colors in each bar indicate mutation signatures. (**B**) Transition of variational lower bound by the number of signatures. (**C**) Predicted mutation signature 1. (**D**) The COSMIC known signatures 2 and 3.

3.3 Novel mutation signatures found in this study

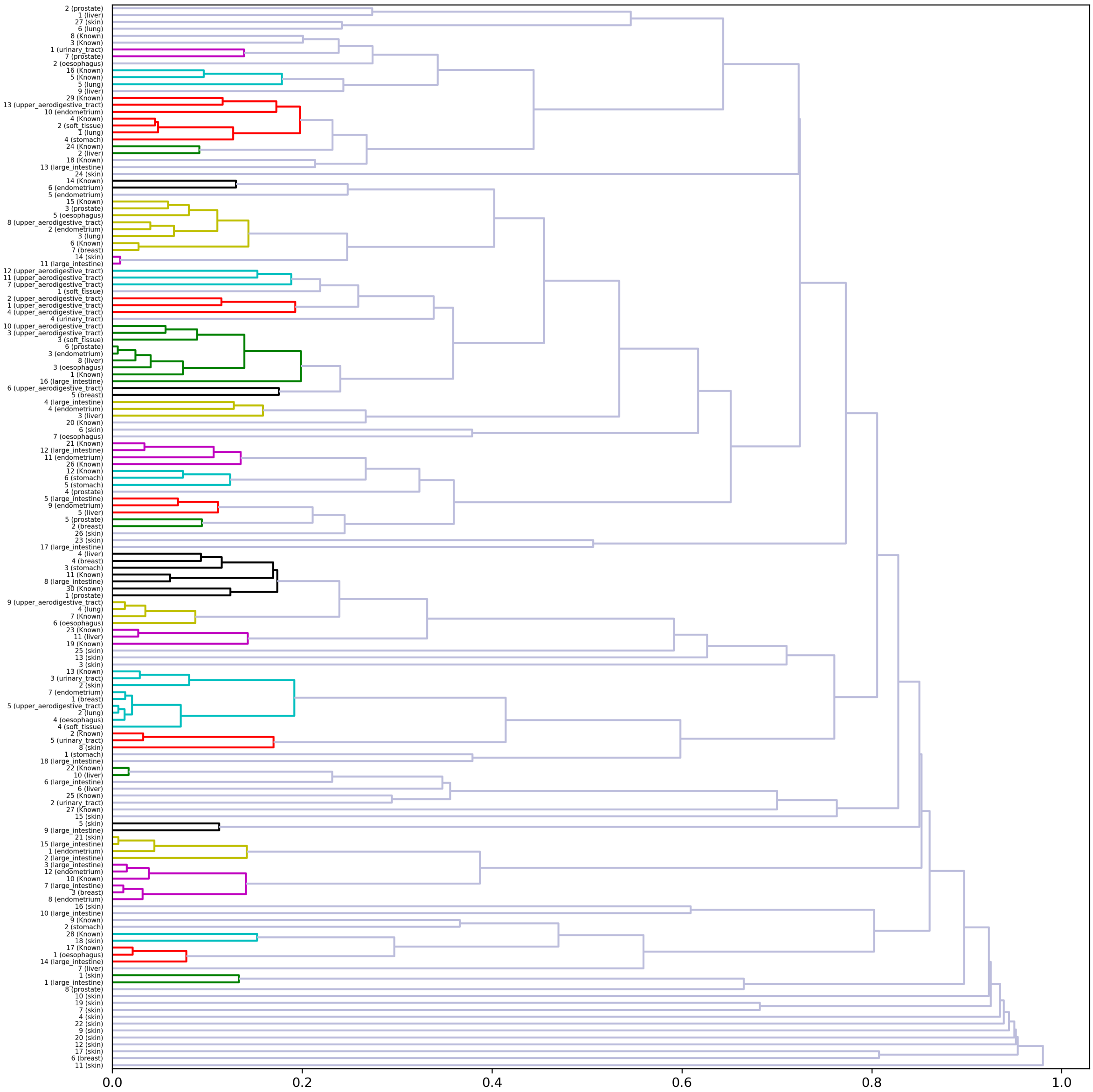


Figure S4: Clustering results for predicted mutation signatures and COSMIC known signatures with the $\mathcal{M}_1$ dictionary.

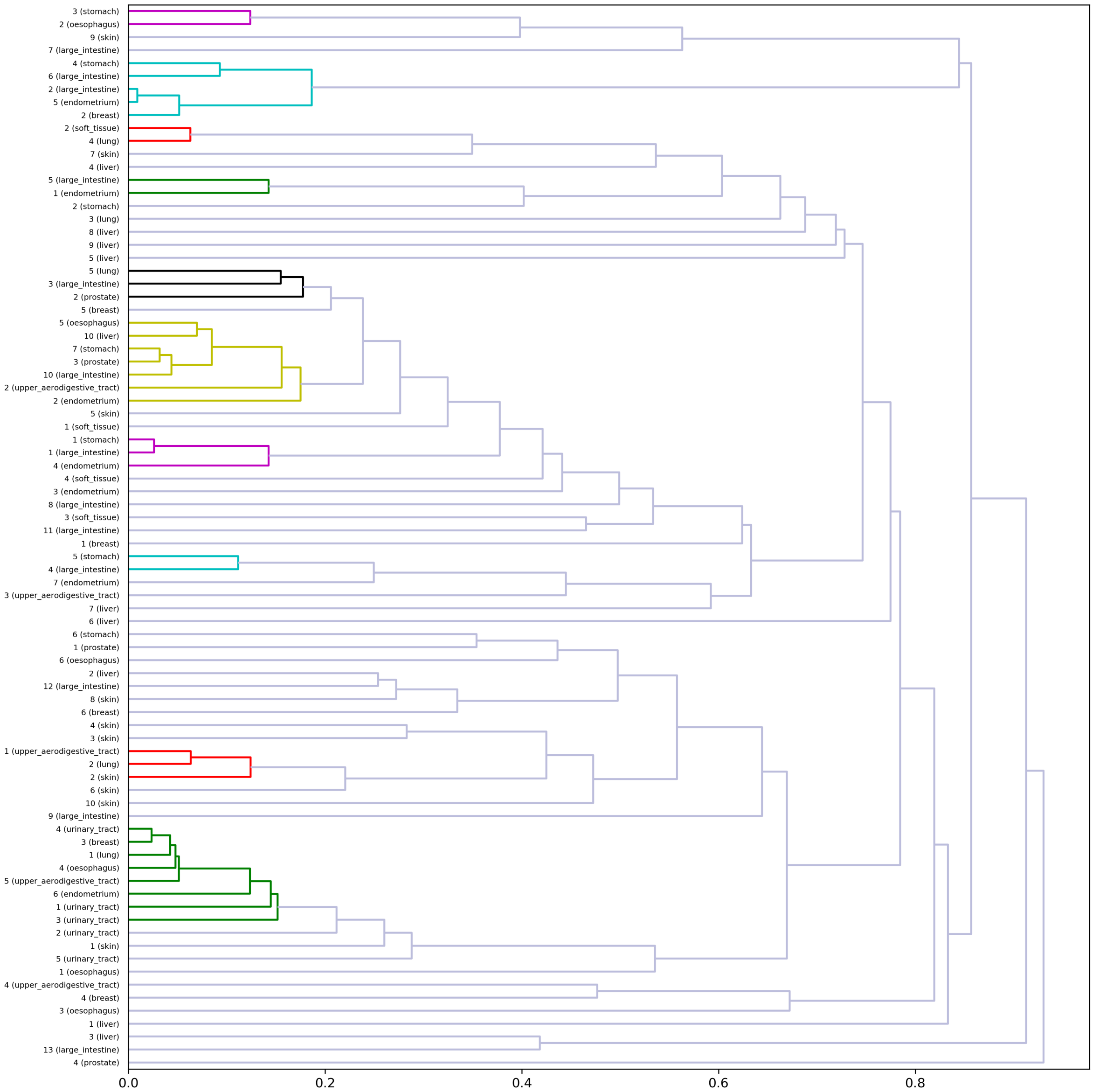


Figure S5: Clustering results for the predicted mutation signatures with the $\mathcal{M}_2$ dictionary.

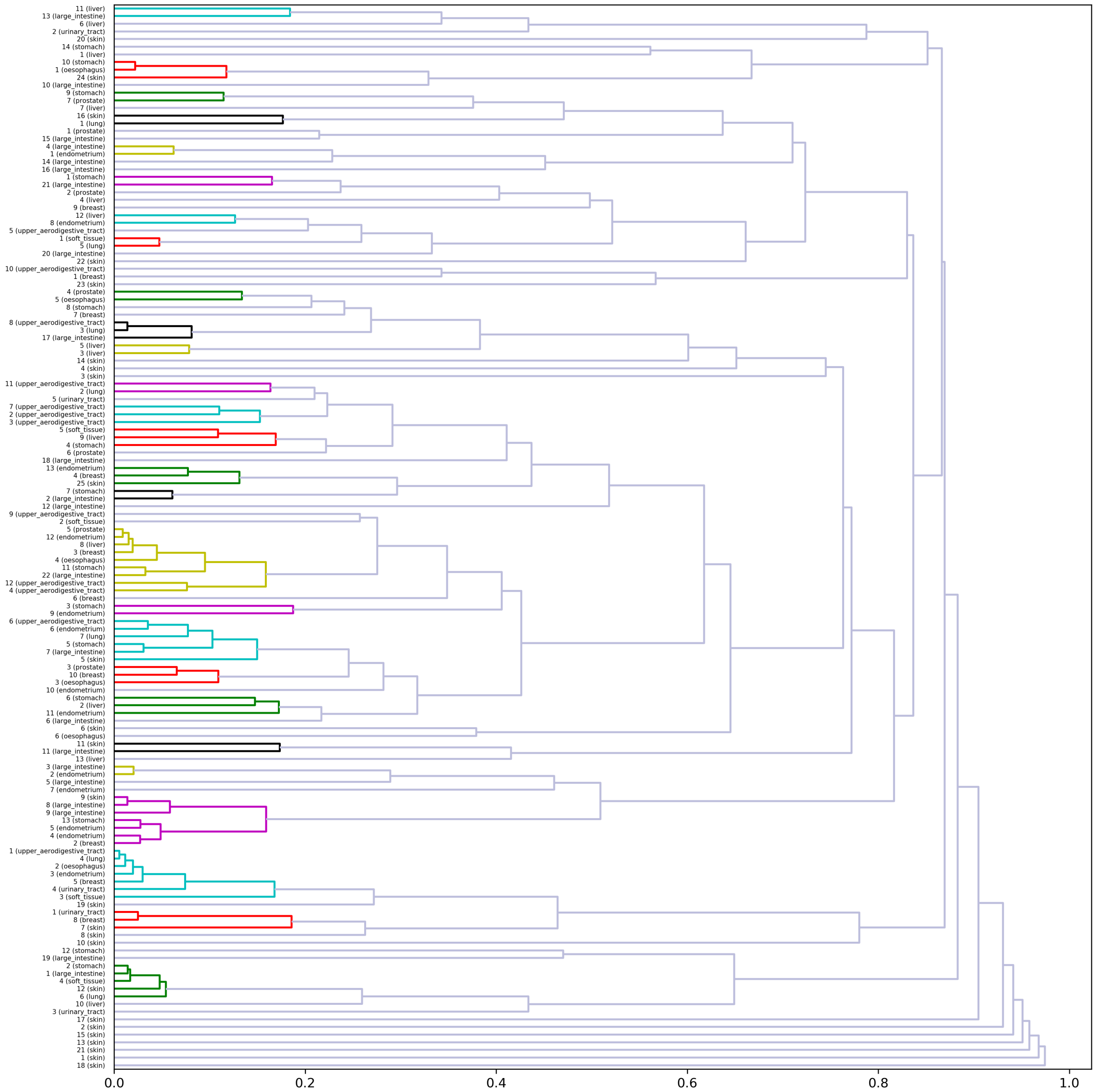


Figure S6: Clustering results for the predicted mutation signatures with the $\mathcal{M}_3$ dictionary.

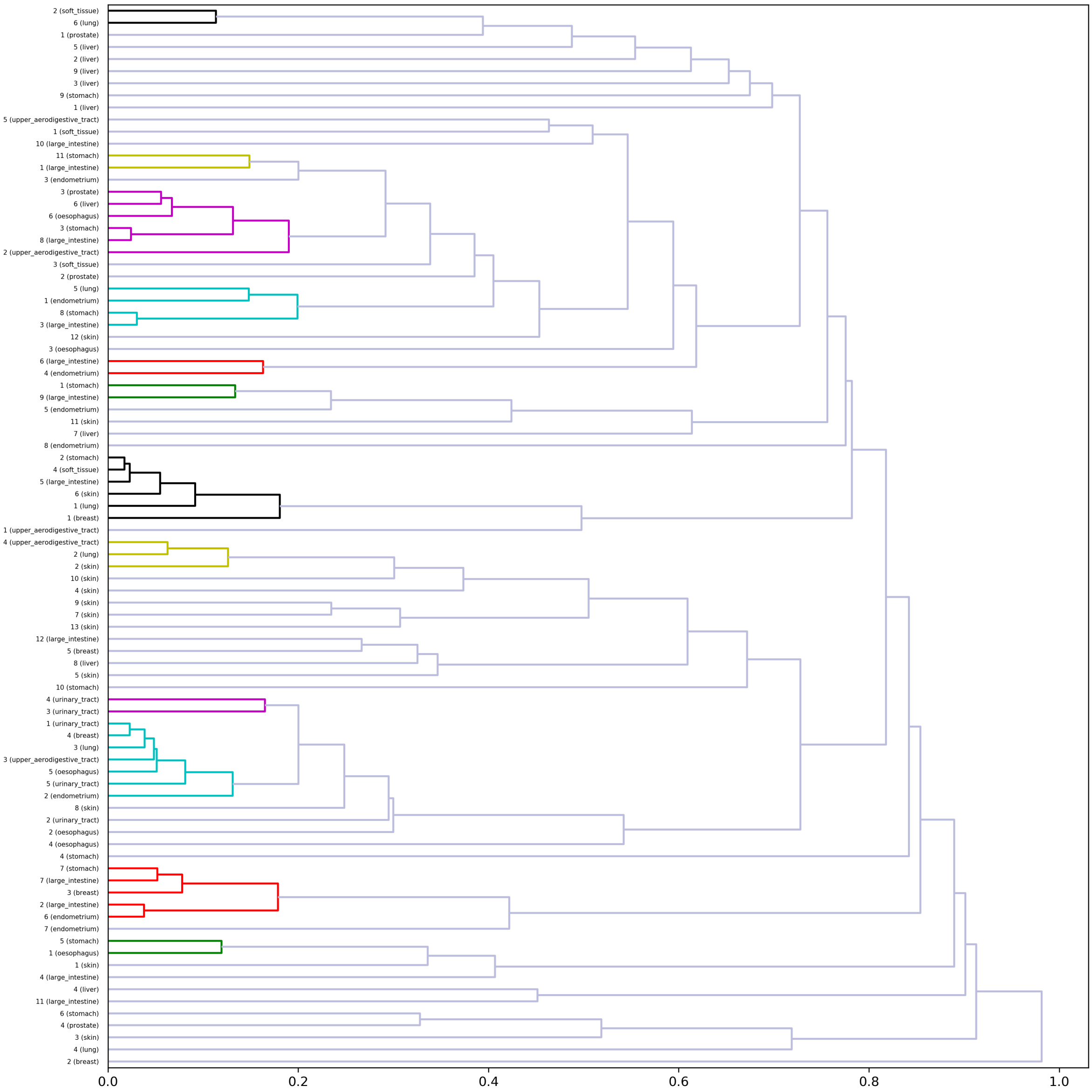


Figure S7: Clustering results for the predicted mutation signatures with the $\mathcal{M}_4$ dictionary.

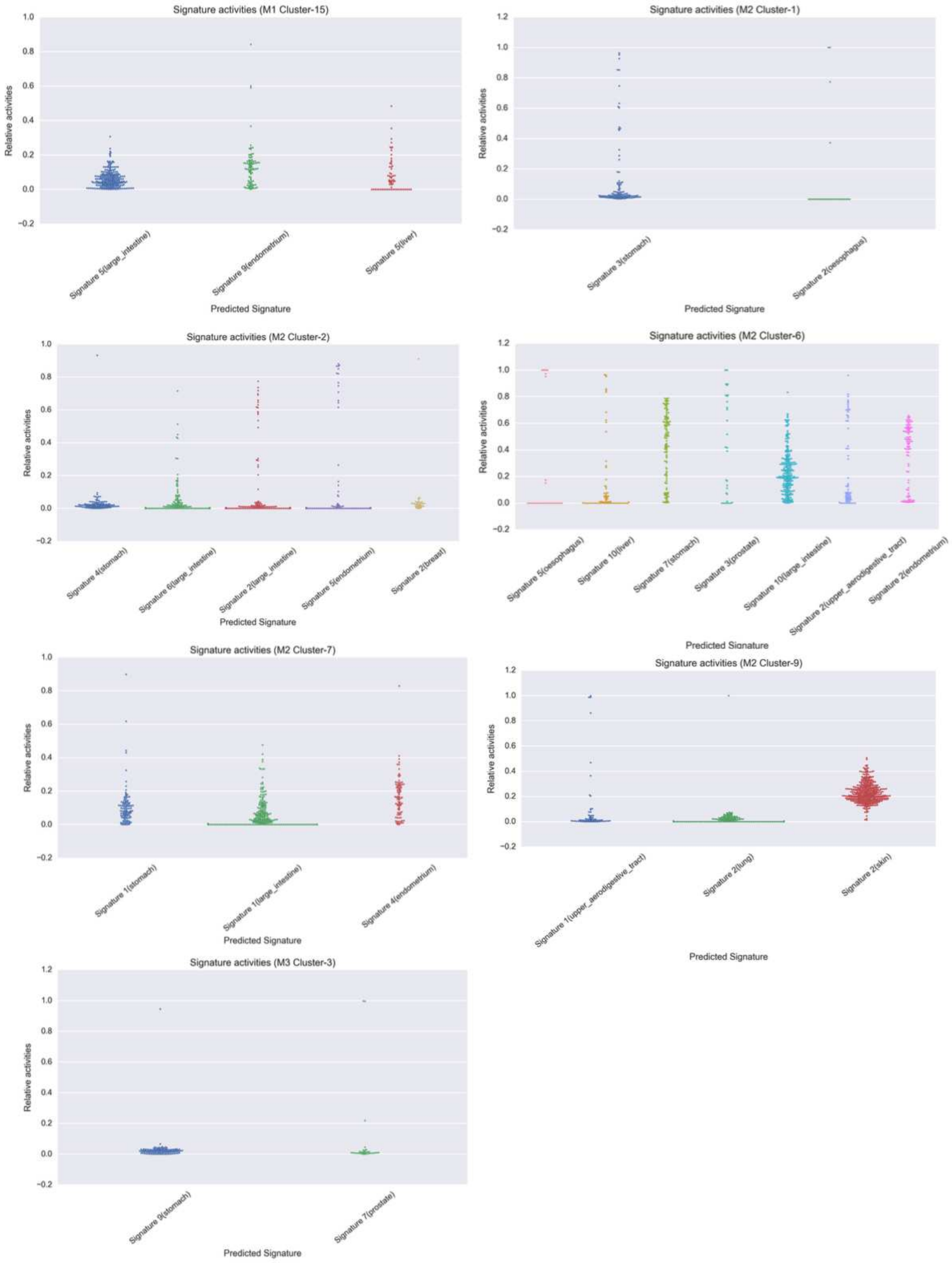


Figure S8: **Signature activities of seven clusters.** Each subfigure shows the signature activities of interesting clusters discussed in the main text. One scatter point corresponds to one sample, and the vertical axis represents relative contribution of the signature (i.e. signature activity).

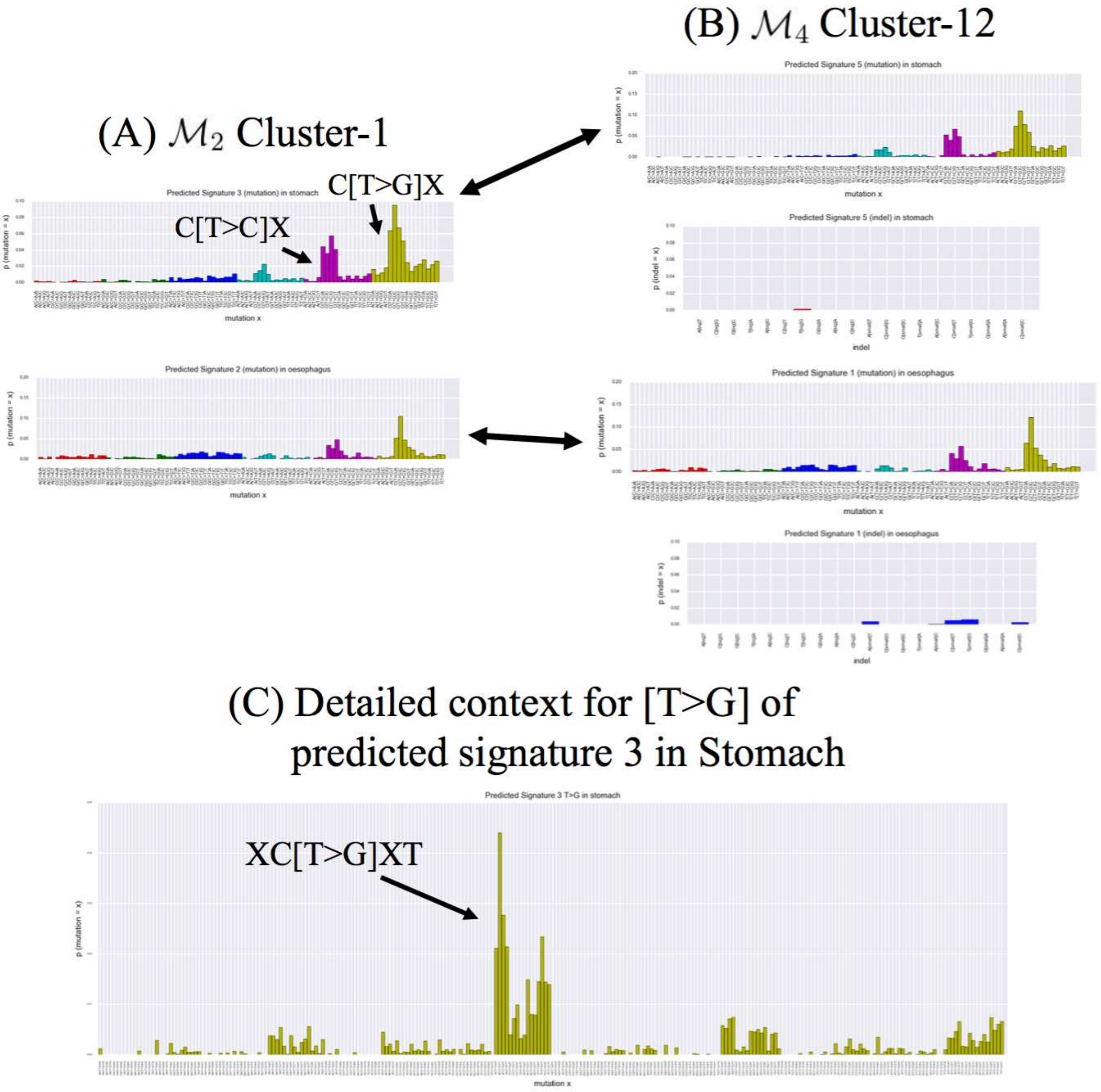


Figure S9: $\mathcal{M}_2$ **Cluster-1 &** $\mathcal{M}_4$ **Cluster-12.** These signatures are found in the stomach and esophagus. Four bar graphs that are coded into 6 colors show the proportion of only a substitution mutation in $\mathcal{M}_2$ and $\mathcal{M}_4$. Mutation contexts of substitutions in $\mathcal{M}_2$ and $\mathcal{M}_4$ are transformed to those of $\mathcal{M}_1$, and the horizontal axis means the type of substitution in $\mathcal{M}_1$. Besides, in the results with $\mathcal{M}_4$ (arranged on the right-hand side), the Indel (insertions and deletions) panel under the substitution panel shows its proportion by each mutation context, so that the horizontal axis means mutation types of $\mathcal{M}_I$ discussed in the main text. (**A**) These signatures have peaks at C[T>C]X and C[T>G]X in common as the arrows indicate. (**B**) Moreover, they have some indels with $\mathcal{M}_4$. (**C**) In addition, they have peaks at XC[T>G]XT when they are analyzed with a mutation context in detail.

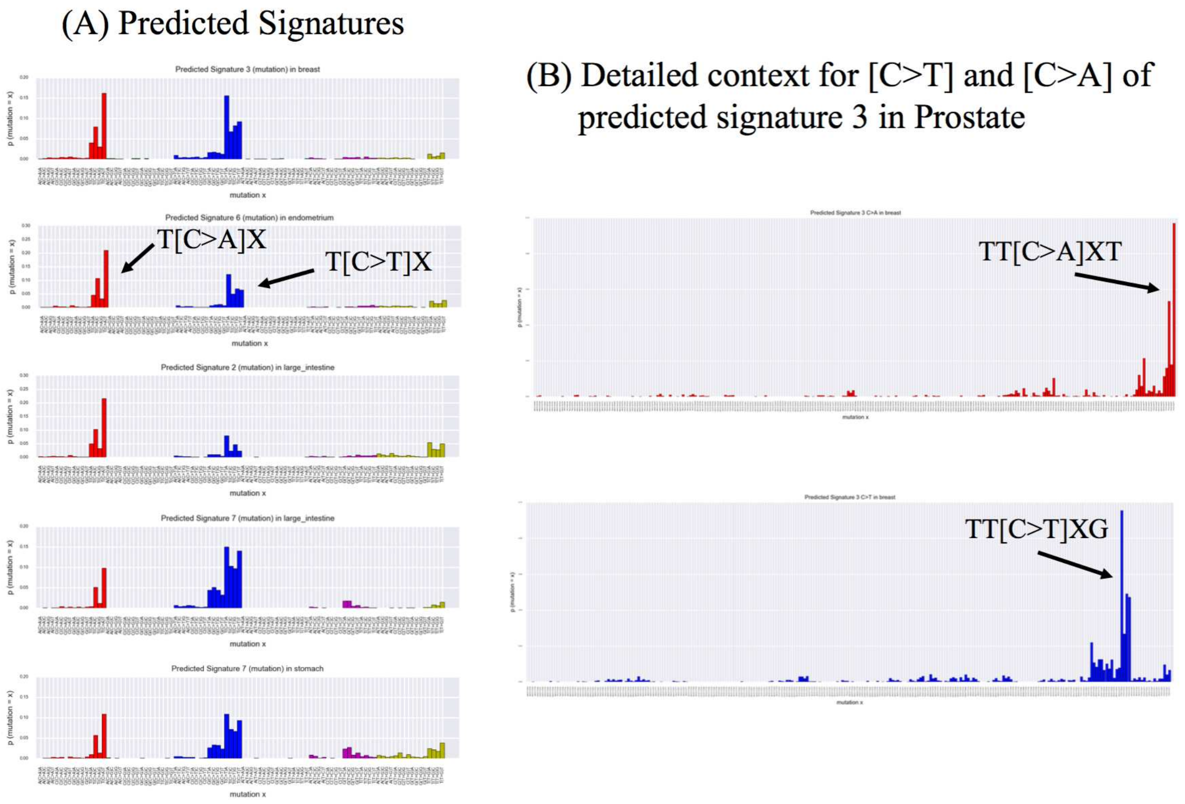


Figure S10: $\mathcal{M}_4$ **Cluster-11.** (**A**) These signatures are found in many organs such as breasts, the endometrium, large intestine, and stomach, then they have peaks at T[C>A]X and T[C>T]X in common as the arrows indicate. (**B**)In the figure showing the context in detail, the horizontal axis means a mutation context of C>A and C>T substitutions in $\mathcal{M}_4$. Readers can see that any signature has large peaks at TT[C>A]XT and TT[C>T]XG particularly.

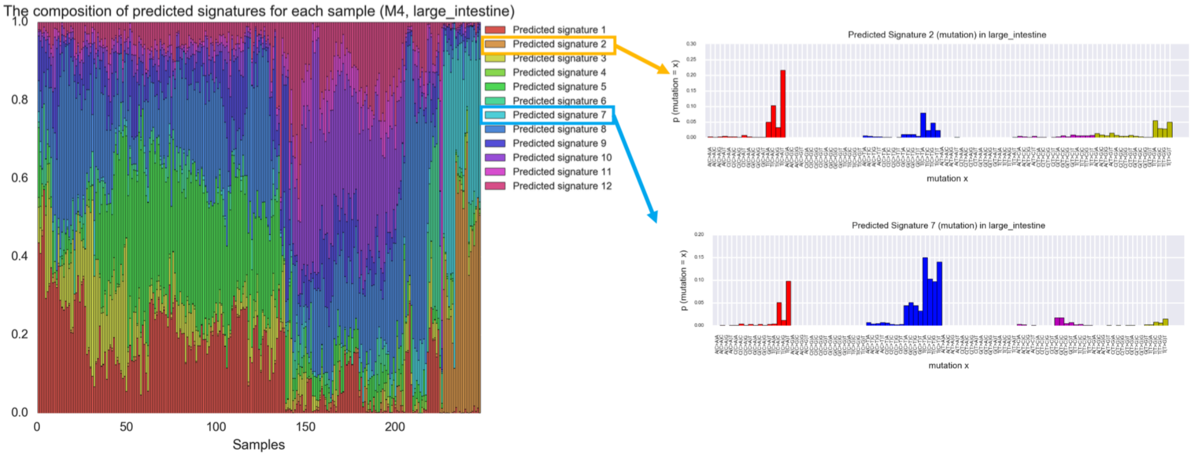


Figure S11: **Signature Activity in large intestine using $\mathcal{M}_4$.** The figure on the left shows the signature activity, the view is similar to Suppl. Fig. S3A. The two figures on the right show the notable signatures contained in $\mathcal{M}_4$ Cluster-11 (Suppl. Fig. S10). The activities of these two signatures are clearly different from each sample, suggesting the existence of a mutational process with multiple kinds of mutational distributions like AID/APOBEC family (i.e. Known Signature 2 and 13).

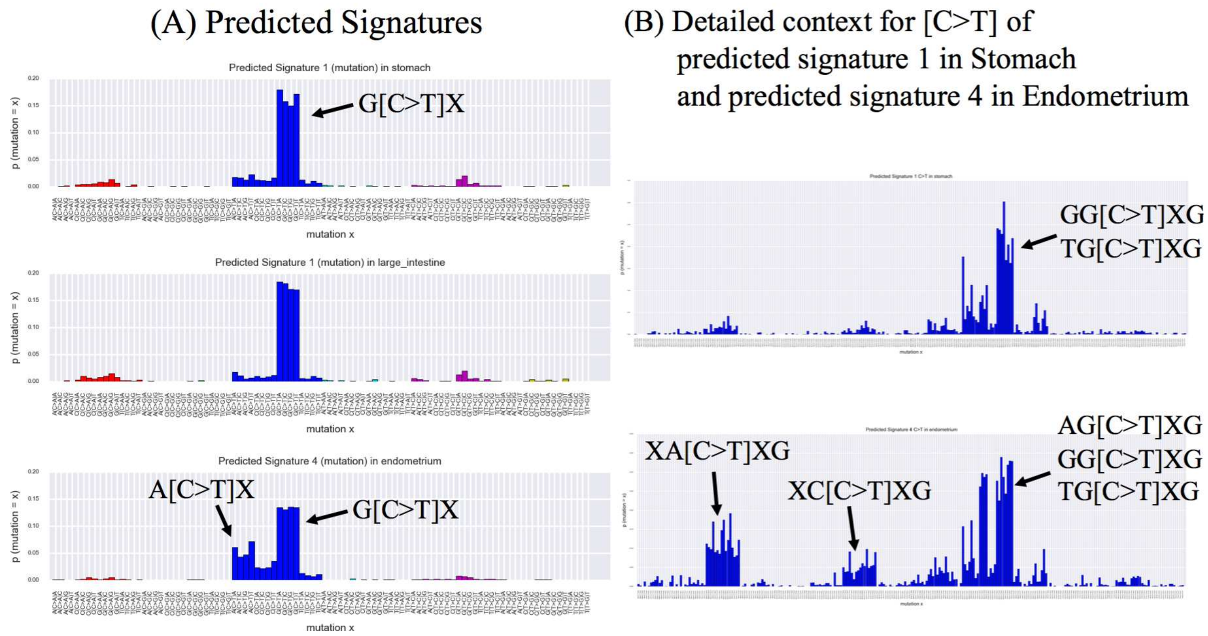


Figure S12: $\mathcal{M}2$ **Cluster-7**. (**A**) These signatures are found in many organs such as the stomach, large intestine, and endometrium; they have peaks at G[C>T]X in common as the arrows indicate. In addition, only the predicted signature in the endometrium has peaks at A[C>T]X. (**B**) When we see the mutation context in detail as is the case for Supplementary Fig. S10, signatures from the stomach and large intestine have large peaks at GG[C>T]XG and TG[C>G]XG particularly.

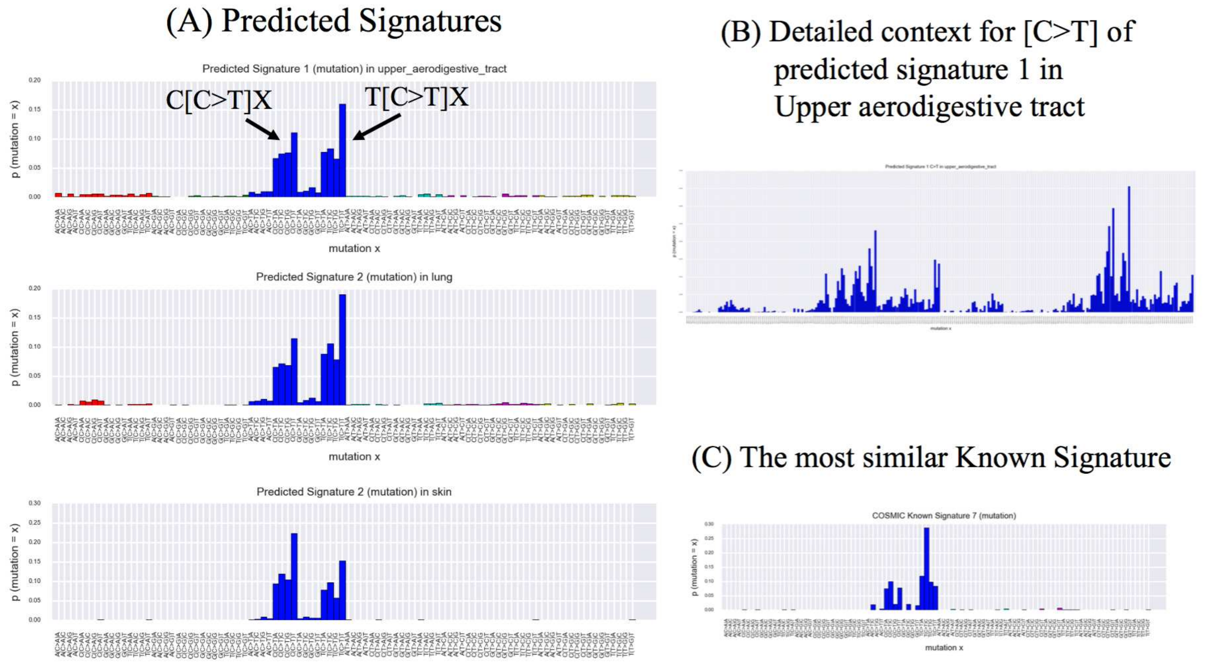


Figure S13: $\mathcal{M}_2$ **Cluster-9**. (**A**) These signatures are found in the upper aerodigestive tract, lungs, and skin; any extracted signature has peaks at C[C>T]X and T[C>T]X as the arrows show. In particular, this tendency is strong when the 3 ′ adjacent base of the mutation is thymine. (**B**) To see mutation context in detail as in Supplementary Fig. S10, there are peaks at XT[C>T]XC. (**C**) The COSMIC Known Signature most similar to these predicted signatures is Signature-7, which is associated with ultraviolet light as its mutational process (e.g., cosine distance between Predicted Signature 2 from lungs with $\mathcal{M}_2$ and COSMIC Known Signature 7 is 0.2076).

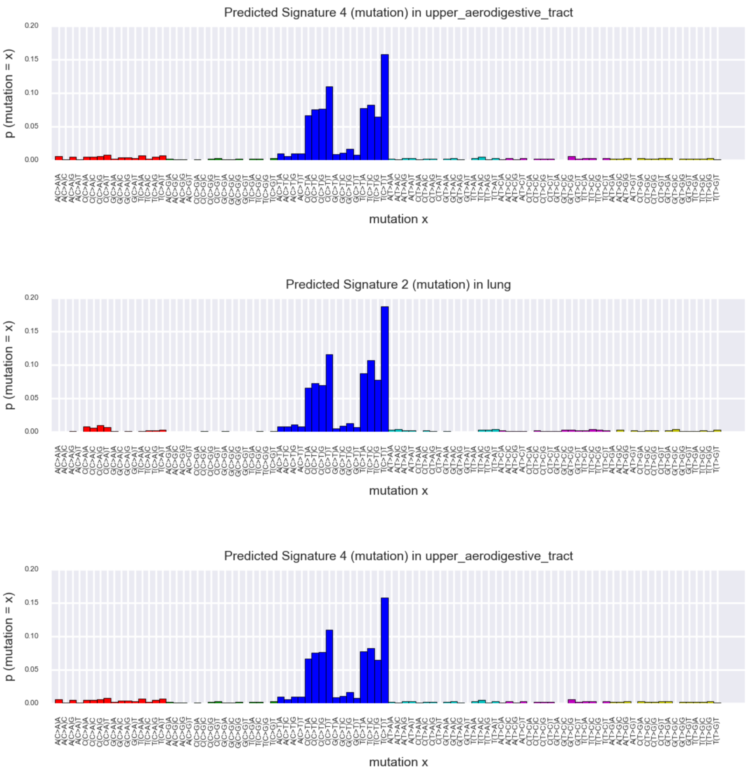


Figure S14: $\mathcal{M}_4$ Cluster-8, where mutational distributions of three predicted signatures are shown.

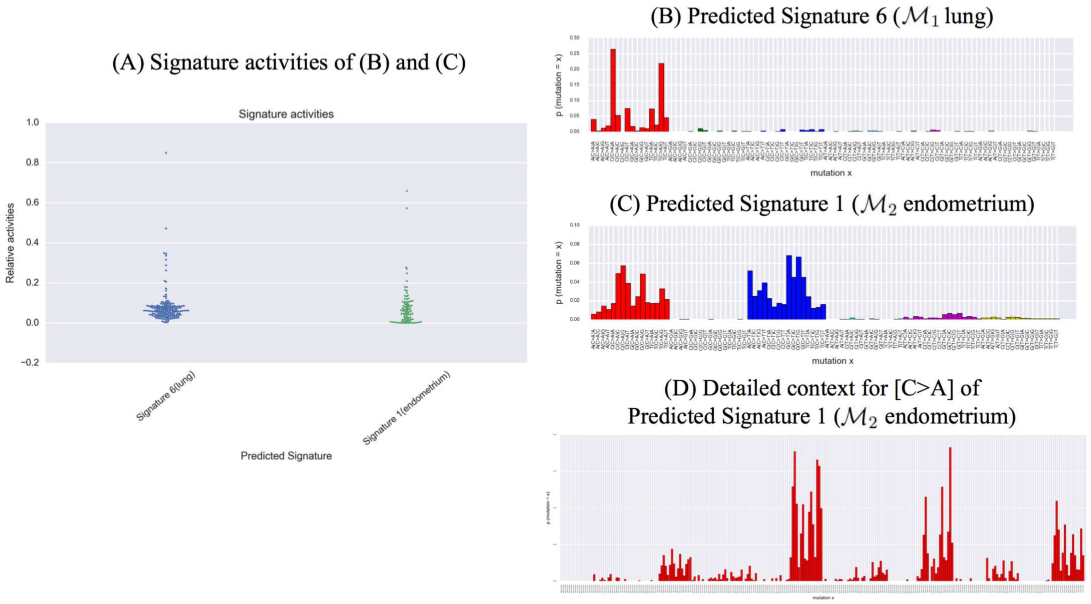
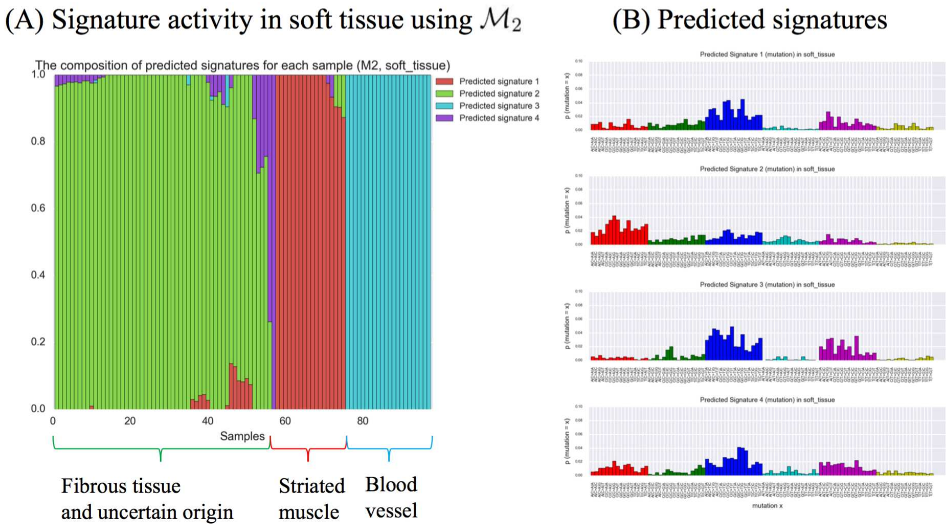


Figure S15: Two predicted mutation signatures (Predicted Signature 6 in lung with $\mathcal{M}_1$ and Predicted Signature 1 in endometrium with $\mathcal{M}_2$). The closest cosine distances to known signatures are more than 0.2. (**A**) Signature activities of the signatures, where both the median and mean values are larger than 0.05. (**B**) Mutational distribution of Predicted Signature 6 in lung with $\mathcal{M}_1$, where peaks at C[C>A]A and T[C>A]G are observed. (**C**) Mutational distribution of Predicted Signature 1 in endometrium with $\mathcal{M}_2$. This signature forms a cluster with Predicted Signature 5 in large intestine with $\mathcal{M}_2$ whose median and mean activities are 0.0413 and 0.0561, respectively. (**D**) The detailed mutational context for C to A mutation of Predicted Signature 1 in endometrium (C), where the peak of XX[C>A]XT appears. Higher resolution figures are available at the online material.



(A) Signature activity in soft tissue using $\mathcal{M}_2$

(B) Predicted signatures

Figure S16: (**A**) Signature activity in soft tissue using $\mathcal{M}_2$ mutation dictionary. (See the caption in Fig. S3A.) In most samples, one of predicted signature 1, 2 and 3 is dominant, which is associated with the type of site subtype information (Fibrous tissue and uncertain origin, striated muscle and blood vessel) in the sample. (**B**) Marginalized mutational distribution (onto the $\mathcal{M}_1$ dictionary) of predicted signatures 1,2,3 and 4 in soft tissue with the $\mathcal{M}_2$ dictionary. Although the mutational distributions of predicted signature 1,3 and 4 look similar, the distributions are different before marginalization (see Supplementary HTML materials); This difference might reflect the features of each site subtype.

4 Comparison with other method

First of all, because EMu performs model selection using BIC as described in Section 1, VB-LDA is expected to be able to estimate the correct number of signatures more accurately in simulation experiments. Therefore, we applied EMu to the simulated dataset (used in Section 3.1) in which VB-LDA estimated the correct number of signatures. In the subsequent experiment, to avoid the local minimum, we allocated the initial value of the parameter 50 times for all the methods and employed the parameter whose BIC value was the best. In the simulation, we prepared two datasets, which are introduced in Section 3.1: "Data1" with the number of mutations in each sample $n_s = 400$ and "Data2" with $n_s = 2000$. Furthermore, for both datasets, the number of samples S was 1000, and hyperparameter α was 0.1. We confirmed that VB-LDA estimated the correct number of signatures (Table 1). Conducting computational experiments under these conditions, we revealed that the predicted number of signatures deviated from the correct number ($K = 10$) when EMu was used (Supplementary Table S5). Nonetheless, as mentioned in the original paper on EMu, setting an appropriate "mutational opportunity" greatly improves the ability to estimate the number of signatures; therefore, when we use simulation data that do not allow EMu to use the mutational opportunity, we cannot claim that VB-LDA is superior to EMu. In addition, the probabilistic model assumed by EMu and that assumed by LDA are different. Therefore, we compared VB-LDA with another model selection by BIC using probabilistic latent semantic analysis (PLSA), which is similar to LDA. PLSA is a model that Dirichlet distribution of a prior is excluded from LDA and estimates the parameters by the EM algorithm and derives BIC (Supplementary Fig. S17). As a result, PLSA estimated K(the predicted number of mutation signatures) = 6 in Data1 where the number of mutations is small ($n_s = 400$) but estimated K as 10 (the correct number) in Data2 where $n_s = 2000$. In addition, all the signatures predicted by PLSA in Data2 were associated one to one with the correct signature set. Accordingly, we believe that it is possible to extract signatures with high precision even for samples with a small number of mutations by model selection using an appropriate prior distribution and variational approximation (used in VB-LDA).

Table S5: A comparison of VB-LDA, EMu, and PLSA on simulated data

Method	Number of signatures (Data1)	Number of signatures (Data2)
VB-LDA	10	10
EMu (uniform)	27	29
PLSA	6	10

Data1 and Data2 are artificial datasets for simulation and were created by means of the generation process of LDA and COSMIC Known Signatures as in Subsection 3.1. In both datasets, the correct number of signatures is 10, the number of samples is 1000, and hyperparameter α is 0.1. In addition, the numbers of mutations for each sample were 400 and 2000 for Data1 and Data2, respectively. VB-LDA was able to estimate the correct number of signatures for both datasets (as also seen in Table 1). Regarding EMu, although correction of a mutational opportunity was not performed, it estimated the signature numbers that deviated from the correct one. In model selection by BIC using PLSA, which is a model similar to LDA, the correct number of signatures was estimated only in Data2.

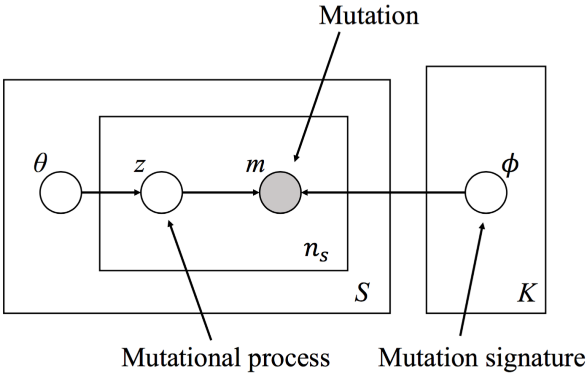


Figure S17: **The graphical model of PLSA.** PLSA is a model where the Dirichlet distribution of the prior is excluded from LDA.

Next, we performed computational experiments with real data. Here, we used $\mathcal{M}_1$ (the number of mutational types is 96) as a mutation dictionary and prepared the breast and lung datasets introduced in VB-LDA analyses (Subsection 3.2.1). We chose EMu as the comparison method, and tested three mutational opportunities: The first one is a uniform distribution (used in the simulation experiment), and the second and third ones were taken from the “human-exome” and “human-genome” files (provided by the authors, which are optimized for each sequencing strategy). According to the original EMu software, a mutation opportunity is created from copy number variations of each sample. Because it is difficult to obtain copy number variation information from COSMIC data, we used such mutation opportunities. For verification of the ability to detect signatures, we focused on how many of each method can extract the signatures that are similar to COSMIC Known Signature, which are confirmed in each primary lesion (whether or not the signatures are similar was determined by whether the cosine distance between them was less than 0.2). The results of estimation of the signature number are listed in Supplementary Table S6A. In the model selection using a mutation opportunity in EMu [see columns “EMu (human-exome)” and “EMu (human-genome)” in the table], small numbers of signatures close to the result of VB-LDA were estimated, but “EMu(uniform)”—where correction was not performed—tended to choose a large number of signatures. A list of predicted signatures similar to COSMIC Known Signatures is described in Supplementary Tables S6B and S6C. EMu could predicted the signatures that are not seen in the result of VB-LDA, and there were also signatures that could be predicted only by VB-LDA. Furthermore, despite not being reported in COSMIC, a signature identified by both EMu and VB-LDA exists (it is Predicted Signature 6 in the breast dataset with $\mathcal{M}_1$, and the cosine distance with the corresponding signature predicted by EMu is $\sim$0.002). These findings are likely to be due to the difference in the method for creating a dataset.

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. Nonetheless, because EMu does not set the mutation opportunity necessary to demonstrate its performance, it is difficult to argue which model is better. In addition, it was difficult to verify the interpretability of the predicted signatures other than by comparison with COSMIC Known Signatures because we do not know the “correct” set of mutation signatures.

Table S6: A comparison of VB-LDA and EMu on real data

(A) The predicted number of signatures

Data	VB-LDA	EMu (human-exome)	EMu (human-genome)	EMu (uniform)
breast ($\mathcal{M}_1$)	7	5	5	9
lung ($\mathcal{M}_1$)	6	4	4	22

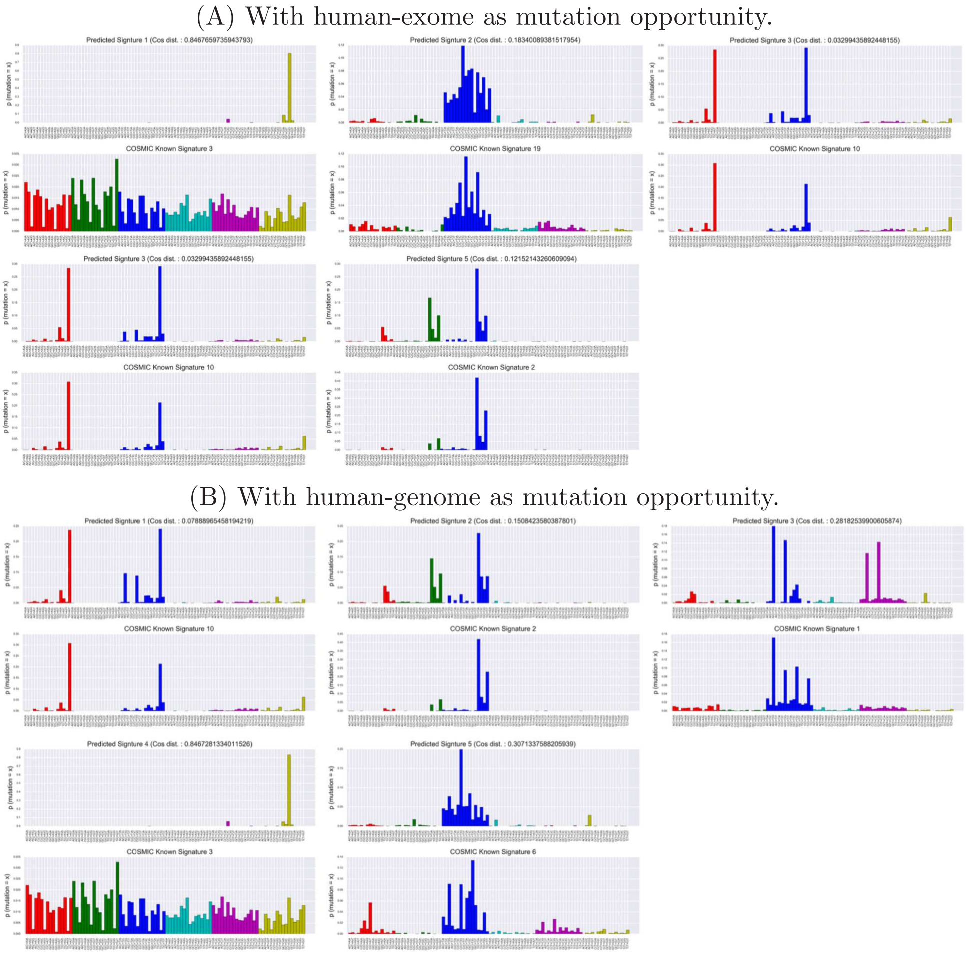
(B) Predicted signatures that are similar to COSMIC Known Signatures seen in the breast dataset

COSMIC Known Signature	1	2	3	5	6	8	10	13	17	18	20	26	30
VB-LDA					○		○	○			○		
EMu (human-exome)		○	○										
EMu (human-genome)		○					○						
EMu (uniform)		○			○		○	○				○	

(C) Predicted signatures that are similar to COSMIC Known Signatures seen in the lung dataset

COSMIC Known Signature	1	2	3	4	5	6	13	15	17
VB-LDA				○	○	○	○		
EMu (human-exome)	○			○					
EMu (human-genome)	○								
EMu (uniform)	○	○			○				

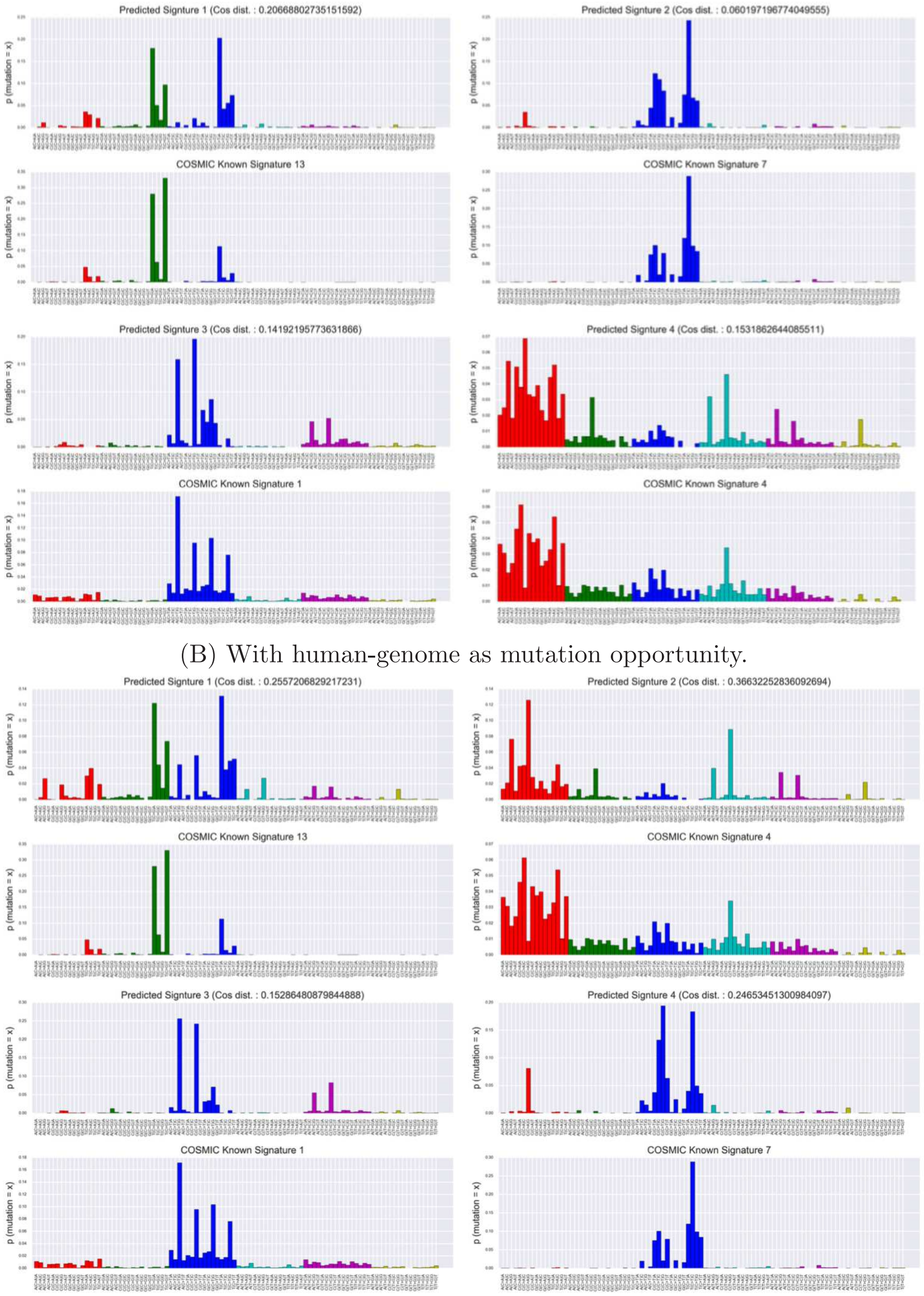
As verification on a real dataset, we used samples of breast and lung tissues in $\mathcal{M}_1$. For EMu, opportunity types are shown in parentheses near each method name; “human-exome” and “human-genome” were prepared in advance and are optimized for each sequencing strategy; “uniform” yields a result when EMu is employed without the correction by opportunity. **(A)** shows the signature number predicted by each method for each dataset. **(B)** and **(C)** present the predicted signatures that are similar to COSMIC Known signatures seen in each primary lesion. The number in the top row represents the list of COSMIC Known Signatures that are considered found in each primary lesion, and the circles in each cell indicate that the predicted signatures that are similar to the corresponding known signatures were extracted (as also confirmed in Supplementary Figs. S18 and S19).



(A) With human-exome as mutation opportunity.

(B) With human-genome as mutation opportunity.

Figure S18: **Signatures predicted by EMu in the breast dataset.**



(A) With human-exome as mutation opportunity.

(B) With human-genome as mutation opportunity.

Figure S19: **Signatures predicted by EMu in the lung dataset.**

5 A new Bayesian hierarchical model for extraction of signatures

We are devising a new model that can extract signatures without separating samples for each primary lesion (Figure S20).

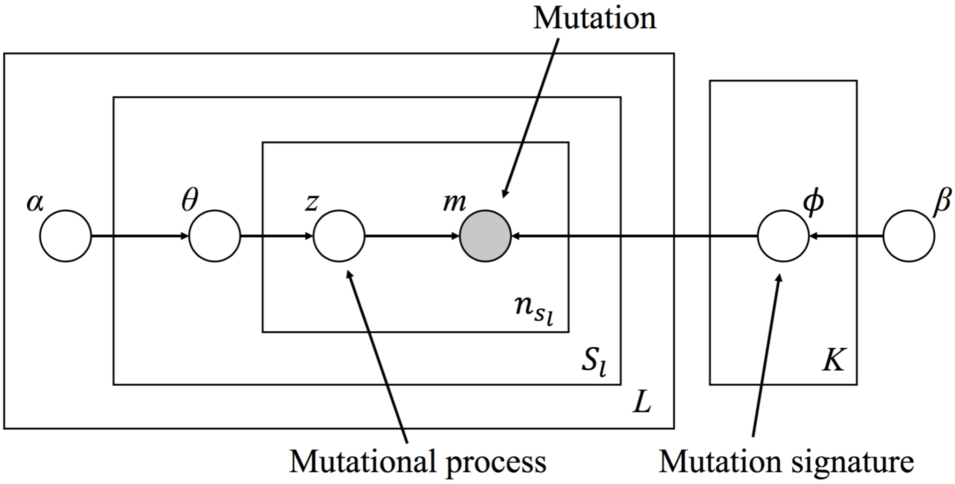


Figure S20: **The new Bayesian hierarchical model.** This model is based on LDA, but the Dirichlet distribution of a prior is hierarchized. L and S_l represent the number of primary lesions and the number of samples for corresponding primary lesion, respectively. The value of α_l differs between primary lesions, so that the model can capture the bias of the activity for mutational process by each primary lesion. In addition, by analyzing all samples at once, researchers can avoid the following problem: mutational distributions of signatures derived from the same mutational process are slightly different among all primary lesions, as seen in other studies.