Supplementary material

Derivation of EM algorithm

The complete log-likelihood including missing data $\{\boldsymbol{z}_i\}$ for the proposed model is

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} I(z_{i,j}=k)\Big(\sum_{l=1}^{L}\log f_{k,l,x_{i,j,l}} + \log q_{i,k}\Big).$$

Here, we introduce the variable for conditional probability for $z_{i,j}$ given the parameters and the mutation features $\boldsymbol{x}_{i,j}$,

$$\theta_{i,k,\boldsymbol{m}} = \Pr\big(z_{i,j}=k \big| \boldsymbol{x}_{i,j}=\boldsymbol{m}, \{\boldsymbol{f}_{k,l}\}, \{\boldsymbol{q}_i\}\big)$$

Note that this conditional probability just depends on the value of mutation feature $\boldsymbol{m} = (m_1, \cdots, m_L)$, not on the index j. Then, the expected complete log-likelihood augmented by Lagrange multipliers is calculated as

$$\sum_{i=1}^{I}\sum_{\boldsymbol{m}} g_{i,\boldsymbol{m}} \sum_{k=1}^{K}\theta_{i,k,\boldsymbol{m}}\Big(\sum_{l=1}^{L}\log f_{k,l,m_l} + \log q_{i,k}\Big) + \sum_{k=1}^{K}\sum_{l=1}^{L}\tau_{k,l}\big(1-\sum_{p=1}^{M_l} f_{k,l,p}\big) + \sum_{i=1}^{I}\rho_i\big(1-\sum_{k=1}^{K} q_{i,k}\big).$$

Differentiating it leads to following stationary equations:

$$\sum_{i=1}^{I}\sum_{\boldsymbol{m}:m_l=p} g_{i,\boldsymbol{m}}\theta_{i,k,\boldsymbol{m}} - \tau_{k,l} f_{k,l,p} = 0,\ (p=1,\cdots,M_l, k=1,\cdots,K,\ l=1,\cdots,L).,$$

$$\sum_{\boldsymbol{m}} g_{i,\boldsymbol{m}}\theta_{i,k,\boldsymbol{m}} - \rho_i q_{i,k} = 0,\ (k=1,\cdots,K, i=1,\cdots,I).$$

Then, by eliminating Lagrange multipliers, updating rules can be obtained.

Relationship with nonnegative matrix factorization

First, for ease of explanation, let assume that the full representation" representation ($L = 1$) is used. Suppose that each $\boldsymbol{m}$ has unique appropriate index from 1 to $|\boldsymbol{M}| = \prod_{l=1}^{L} M_l$ (the number of possible mutation patterns), so that $\boldsymbol{m}$ can be indices of matrices.

Let $G = \{g_{i,\boldsymbol{m}}\}$ denote the $I \times |\boldsymbol{M}|$ matrix, where $g_{i,\boldsymbol{m}}$ is the number of mutations whose mutation patters are equal to $\boldsymbol{m}$ in the i-th cancer genome. Nonnegative matrix factorization aims to find low rank decomposition, $G \sim \tilde{Q}F$, where $\tilde{Q} = \{\tilde{q}_{i,k}\}$ and $F = \{f_{k,\boldsymbol{m}}\}$ are nonnegative matrix, and row vectors of F are often restricted to be sum to one. We used the notation $\tilde{Q}$ instead of Q to represent that the row vectors of $\tilde{Q}$ are not normalized to sum to one in general.

For solving NMF, the previous study (Lee et al. 2000) used the following updating rule:

$$f_{k,m} \leftarrow f_{k,m}\frac{(\tilde{Q}^T G)_{k,m}}{(\tilde{Q}^T\tilde{Q}F)_{k,m}},\quad \tilde{q}_{i,k} \leftarrow \tilde{q}_{i,k}\frac{(GF^T)_{i,k}}{(\tilde{Q}FF^T)_{i,k}},$$

that reduces the *Euclidean distance* $||G - \tilde{Q}F||$. Therefore, the optimization problem for the existing approach is

$$\begin{aligned}
&\text{minimize} && ||G-\tilde{Q}F|| \\
&\text{subject to} && \sum_{\boldsymbol{m}} f_{k,\boldsymbol{m}} = 1,\ k=1,\cdots,K \\
& && f_{k,\boldsymbol{m}} \geq 0,\ k=1,\cdots,K,\ \boldsymbol{m}\in M \\
& && \tilde{q}_{i,k} \geq 0,\ i=1,\cdots,I,\ k=1,\cdots,K.
\end{aligned} \tag{1}$$

On the other hand, there is another type of updating rule:

$$f_{k,m} \leftarrow f_{k,m}\frac{\sum_i \tilde{q}_{i,k} g_{i,m}/(\tilde{Q}F)_{i,m}}{\sum_i \tilde{q}_{i,k}},$$

$$\tilde{q}_{i,k} \leftarrow \tilde{q}_{i,k}\frac{\sum_m f_{k,m} g_{i,m}/(\tilde{Q}F)_{i,m}}{\sum_m f_{k,m}}.$$

that reduces the Kullback-Liebler Divergence:

$$KL(G||\tilde{Q}F) = \sum_{i,m}\Big(g_{i,m}\log\frac{g_{i,m}}{(\tilde{Q}F)_{i,m}} - g_{i,m} + (\tilde{Q}f)_{i,m}\Big).$$

In general cases including the independent representation, there is restrictions $f_{k,\boldsymbol{m}} = \prod_l f_{k,l,m_l}$ by smaller set of parameters. Let us consider the following optimization problem with the Kullback-Liebler Divergence and the restrictions on F:

$$\begin{aligned}
\text{minimize} \quad & KL(G||\tilde{Q}F) \\
\text{subject to} \quad & f_{k,\boldsymbol{m}} = \prod_l f_{k,l,m_l},\ k = 1, \cdots, K,\ \boldsymbol{m} \in M \\
& f_{k,l,p} \geq 0,\ k = 1, \cdots, K,\ \boldsymbol{m} \in M \\
& \tilde{q}_{i,k} \geq 0,\ i = 1, \cdots, I,\ k = 1, \cdots, K.
\end{aligned} \tag{2}$$

In fact, this is equivalent to the proposed method, whose optimization problem can be written as:

$$\begin{aligned}
\text{maximize} \quad & L(Q, F|G)\big(= \sum_{i,m} g_{i,m}\log(QF)_{i,m}\big) \\
\text{subject to} \quad & f_{k,\boldsymbol{m}} = \prod_l f_{k,l,m_l},\ k = 1, \cdots, K,\ \boldsymbol{m} \in M \\
& f_{k,l,p} \geq 0,\ k = 1, \cdots, K,\ \boldsymbol{m} \in M \\
& \sum_k q_{i,k} = 1,\ i = 1, \cdots, I \\
& q_{i,k} \geq 0,\ i = 1, \cdots, I,\ k = 1, \cdots, K.
\end{aligned} \tag{3}$$

Proposition 1 *When $(Q, F) = (Q^*, F^*)$ is an optimal solution of the optimization problem (3), then $(\tilde{Q}, F) = (R^*Q^*, F^*)$ is an optimal solution of the optimization problem (2). On the other hand, when $(\tilde{Q}, F) = (\tilde{Q}^*, F^*)$ is an optimal solution of the optimization problem (2), then $(Q, F) = (R^{*-1}\tilde{Q}^*, F^*)$ is an optimal solution of the optimization problem (3), where $R^* = diag(r_1^*, \cdots, r_I^*), r_i^* = \sum_{\boldsymbol{m}} g_{i,\boldsymbol{m}}, i = 1, \cdots, I$.*

Proof. This is because

$$KL(G||\tilde{Q}F) = -\sum_i\Big((\sum_m g_{i,m})\log\tilde{r}_i - \tilde{r}_i\Big) - L(Q, F|G) + (\text{constant value}),$$

where Q is row-normalized matrix for $\tilde{Q}$, $\tilde{r}_i = \sum_k q_{i,k}$ for each i, and $(\sum_m g_{i,m})\log\tilde{r}_i - \tilde{r}_i$ takes its maximum at $\tilde{r}_i = r_i^*$. □

Table 1: The number of mutation signatures selected for each cancer type.

cancer type	# mutation signature
AML	2
ALL	3
Bladder	3
Breast	4
Cervix	3
CLL	3
Colorectum	5
Esophageal	4
Glioblastoma	3
Glioma-Low-Grade	3
Head-and-Neck	5
Kidney-Chromophobe	3
Kidney-Clear-Cell	3
Kidney-Papillary	3
Liver	3
Lung-Adeno	4
Lung-Small-Cell	4
Lung-Squamous	3
Lymphoma-B-Cell	4
Medulloblastoma	3
Melanoma	4
Myeloma	3
Neuroblastoma	3
Ovary	3
Pancreas	4
Pilocytic-Astrocytoma	4
Prostate	2
Stomach	5
Thyroid	4
Uterus	4

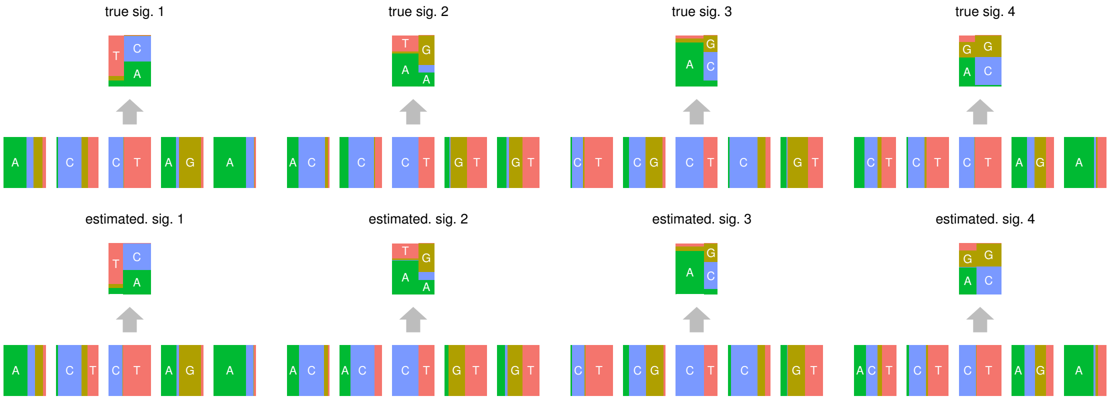


Figure 1: An example of the relationship between true and estimated mutation signatures in the simulation. In this example, the numbers of cancer genomes and mutations for each cancer genome are 25 and 100, respectively, and the parameters α and γ are both set to 1. From this figure, we can see that fairly accurate estimation is possible even with moderate numbers of cancer genomes and mutations

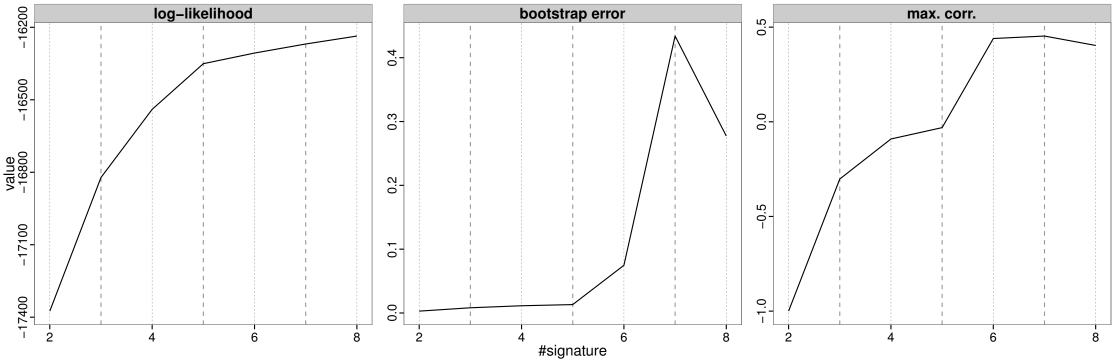
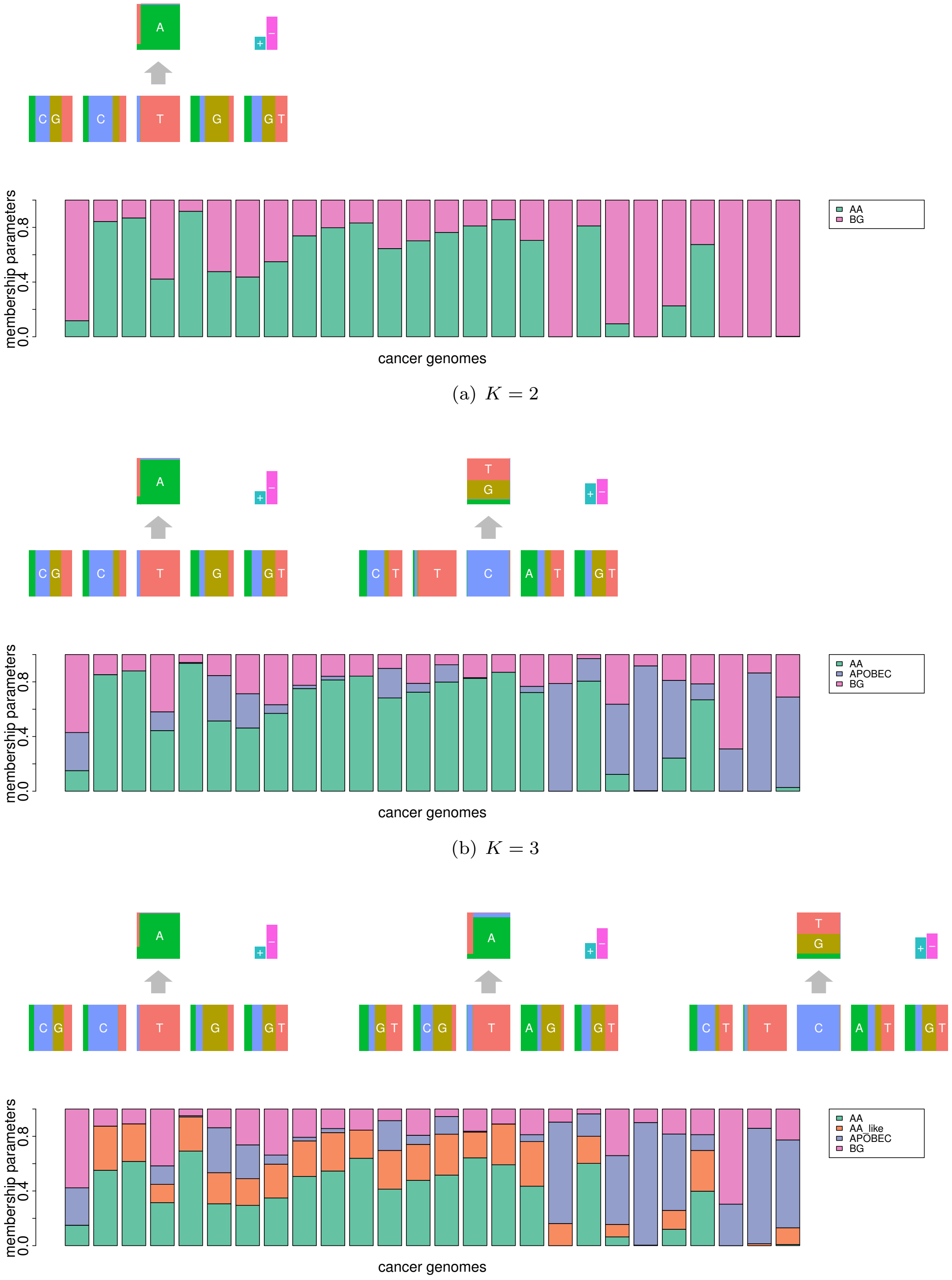


Figure 2: The log-likelihood, bootstrap-errors and maximum correlation values among estimated membership parameters for several numbers of signatures K in the trial of the Supplementary Figure 1.



(a) $K = 2$

(b) $K = 3$

(c) $K = 4$

Figure 3: The result of estimated mutation signatures and membership parameters for UTUC data when changing the number of mutation signatures K.

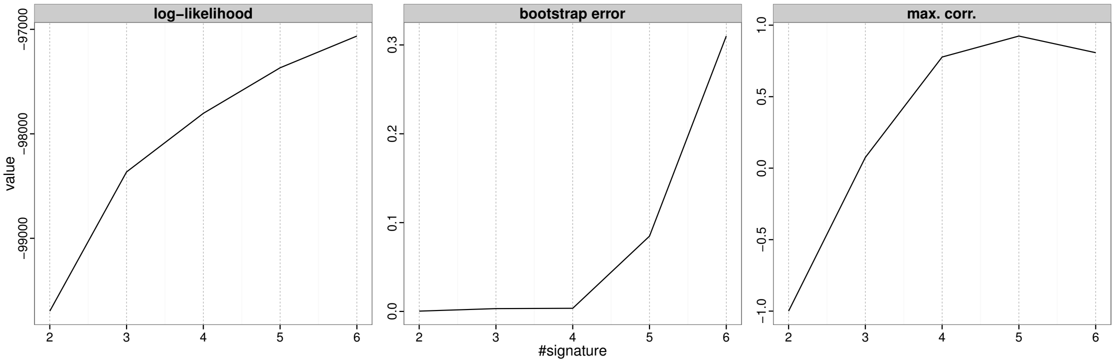


Figure 4: The log-likelihood, bootstrap-errors and maximum correlation values among estimated membership parameters for several numbers of signatures K in UTUC data.

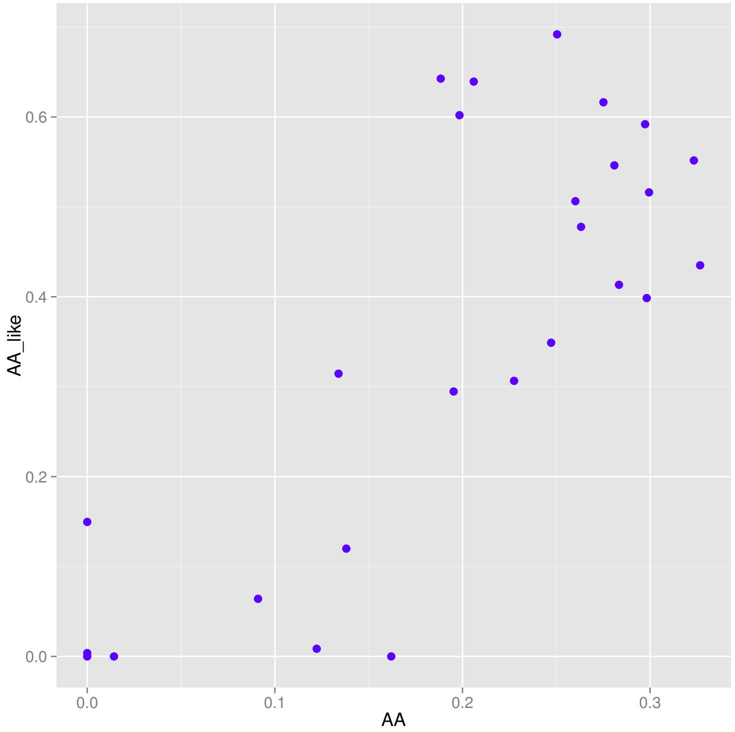


Figure 5: The relationships between the two membership parameters, AA (the first signature in the Supplementary Figure 3) and AA_like (the second signature in the Supplementary Figure 3) signatures.

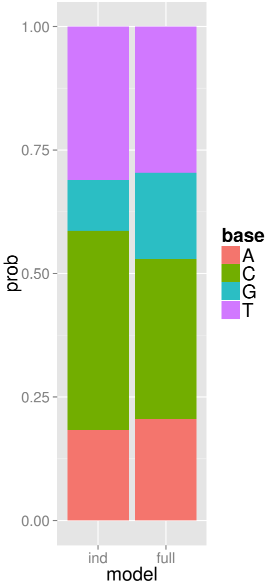
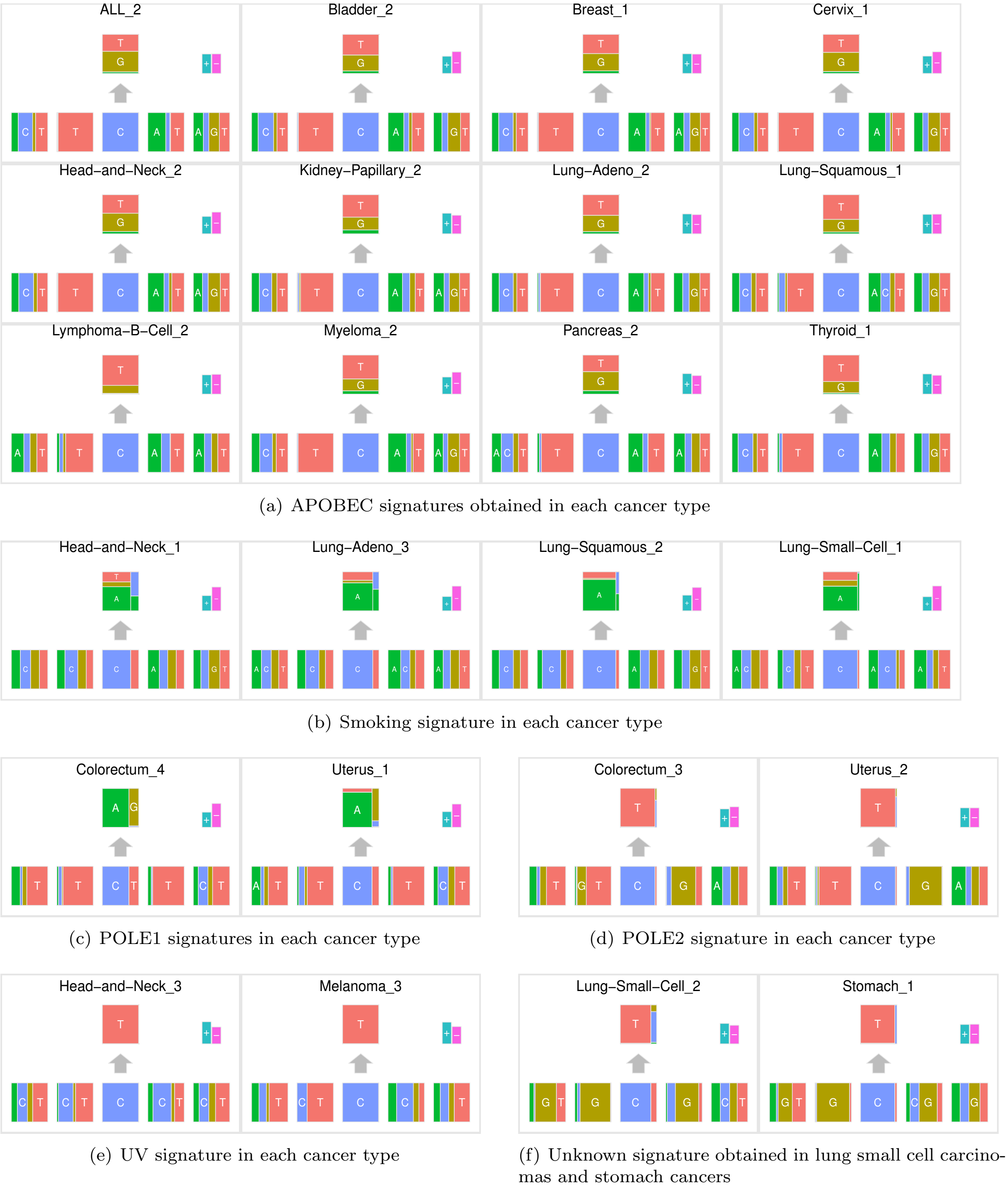


Figure 6: The frequencies of bases at two 5' to the mutated site for the APOBEC mutation signatures obtained in UTUC data using the independent and full models.



(a) APOBEC signatures obtained in each cancer type

(b) Smoking signature in each cancer type

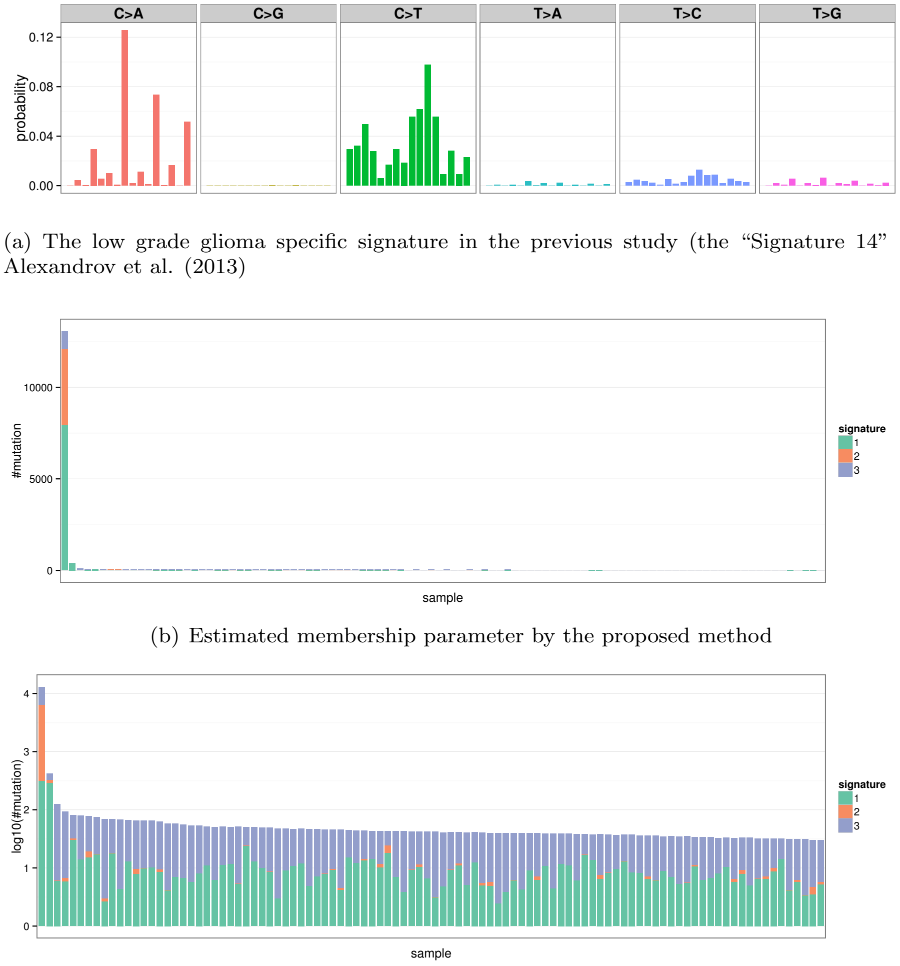
(c) POLE1 signatures in each cancer type

(d) POLE2 signature in each cancer type

(e) UV signature in each cancer type

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

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

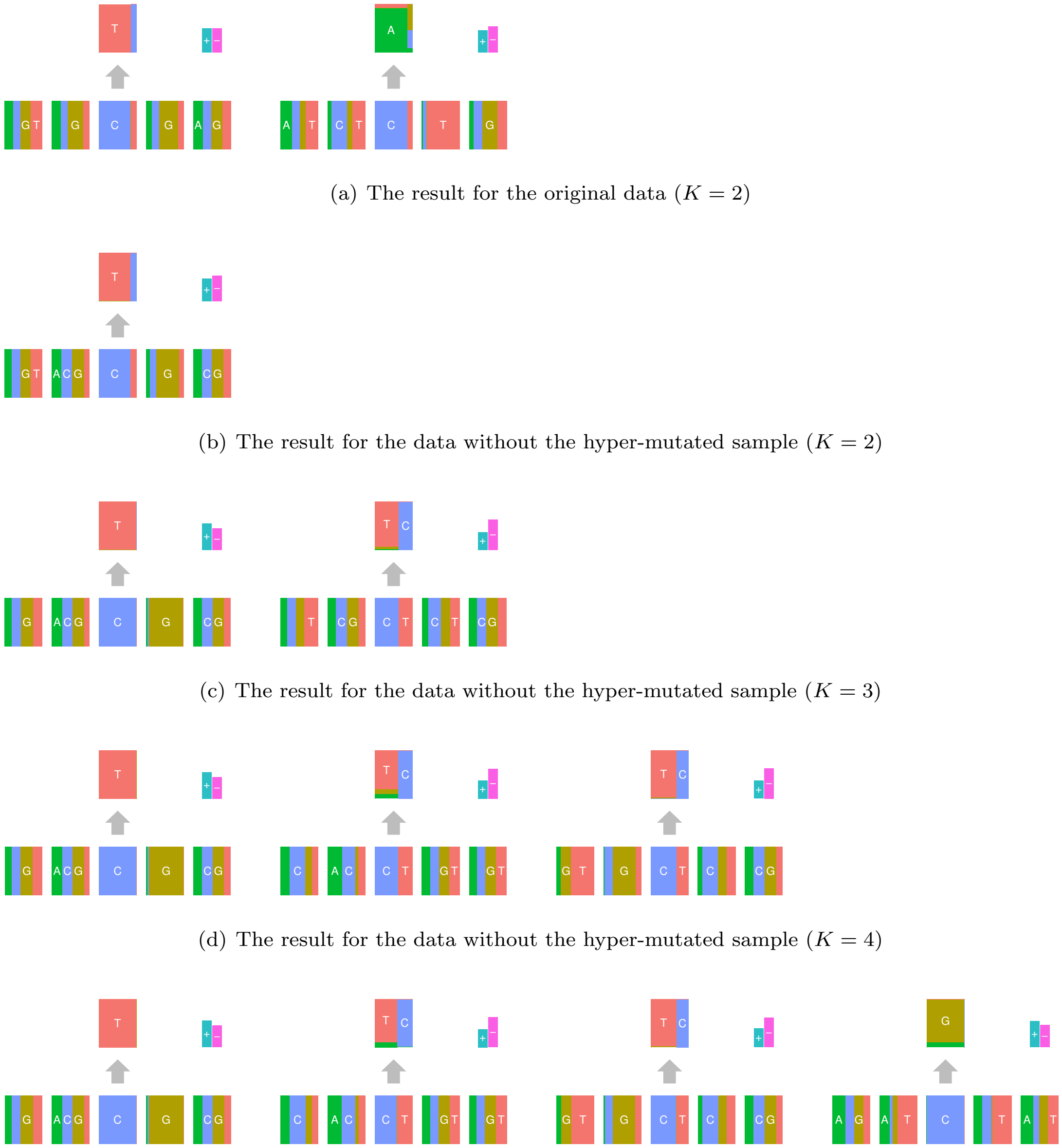


(a) The low grade glioma specific signature in the previous study (the "Signature 14" in Alexandrov et al. (2013)

(b) Estimated membership parameter by the proposed method

(c) Estimated membership parameter by the proposed method in log scale

Figure 8: (a) The barplot is divided by 6 substitution patterns. In each division, 16 bars show joint probabilities of 16 combinations of the immediate 5' and 3' bases. (b, c) We have selected top 100 cancer samples according to the number of mutation. The height of bar shows (the logarithm of) the number of mutations for each sample, and the ratio of colored division shows the ratio of estimated membership parameters for each signature and sample. The low grade glioma specific signature detected by the proposed method is the signature 2. We can see that the mutations corresponding to signature 2 is mostly from the sample with an extremely high mutation rate.



(a) The result for the original data ($K = 2$)

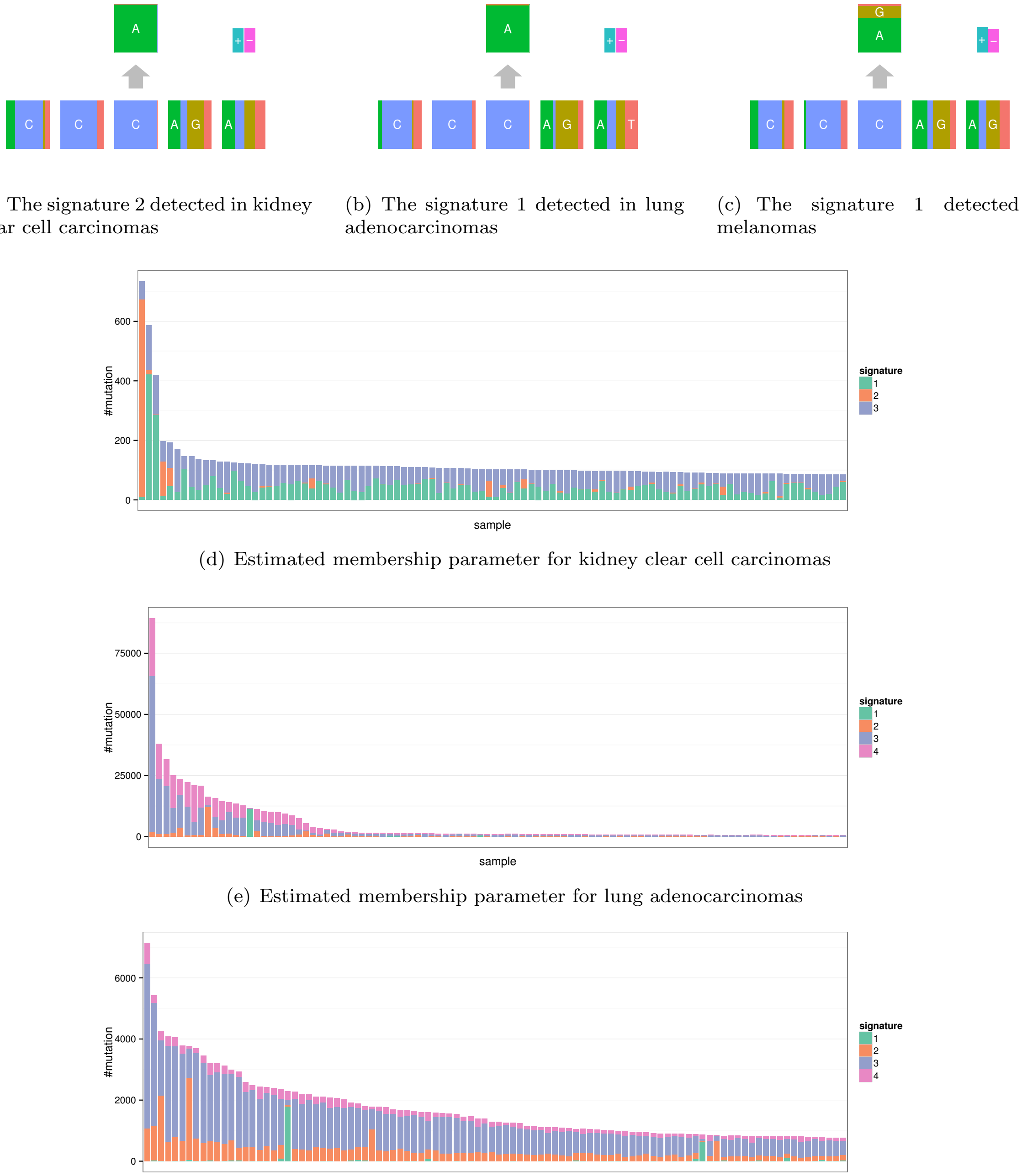
(b) The result for the data without the hyper-mutated sample ($K = 2$)

(c) The result for the data without the hyper-mutated sample ($K = 3$)

(d) The result for the data without the hyper-mutated sample ($K = 4$)

(e) The result for the data without the hyper-mutated sample ($K = 5$)

Figure 9: (a) The signatures obtained for the original data. The first (from the left) signature seems to be deamination of 5-methyl-cytosine. The second signature is the low grade glioma signature. (b, c, d, e) The signature obtained for the data without the hyper-mutated case for several number of mutation signatures (K, including the background signature). Although the signature related to deamination of 5-methyl-cytosine remained, low grade glioma specific signature could not be observed.



(a) The signature 2 detected in kidney clear cell carcinomas

(b) The signature 1 detected in lung adenocarcinomas

(c) The signature 1 detected in melanomas

(d) Estimated membership parameter for kidney clear cell carcinomas

(e) Estimated membership parameter for lung adenocarcinomas

(f) Estimated membership parameter for melanomas

Figure 10: (a, b, c) The putative oxidative artifact signatures estimated for each cancer. We can observe consistent abundance of the base C at the −2 position. (d, e, f) For each cancer type, We have selected top 100 cancer samples according to the number of mutation. The height of bar shows the number of mutations for each sample, and the ratio of colored division shows the ratio of estimated membership parameters for each signature and sample. We can see that the signature corresponding to putative oxidative artifacts concentrates on a small number of samples.