Appendix

A Sufficient Condition for Improving Bayes Factor

In the main content of our paper, we try to construct the generative model, which satisfies the following condition.

Condition 1 (Improvement Condition for Bayes factor)**.**

Bayes factor is improved if

$$\sum_{m\in\{-1,1\}} P(m)\cdot m\cdot \sum_{\omega\in\Omega}(\ln B_{new}(\omega)-\ln B(\omega))P_m(\{\omega\})\geq 0, \tag{1}$$

$$m=\begin{cases} 1 & \text{if mutation exists} \\ -1 & \text{if mutation does not exist} \end{cases}$$

As in the main content of our paper, it is enough to check the following condition.

Condition 2 (Sufficient condition for improvement of Bayes factor)**.**

Bayes factor is improved if

$$\begin{aligned} &\forall t\in\{0,...,T-1\},\forall m\in\{-1,1\}, \\ &\text{such that } m\cdot\sum_{d_t\in S_t}(\ln b_{new}(d_t)-\ln b(d_t))P_m(\{\omega|D_t(\omega)=d_t\})\geq 0 \end{aligned} \tag{2}$$

Using the real data sets, we checked the condition of Eq. (2). Generative models of OVarCall and HapMuC are used for computing Bayes factor.

A.1 Improvement on Bayes Factor when Heterozygous SNPs Information Is Available

Using The Cancer Genome Atlas (TCGA) Mutation Calling Benchmark 4 datasets, we confirmed the sufficient condition, when heterozygous SNPs information is available. HapMuC generative model is used for computing Bayes factor.

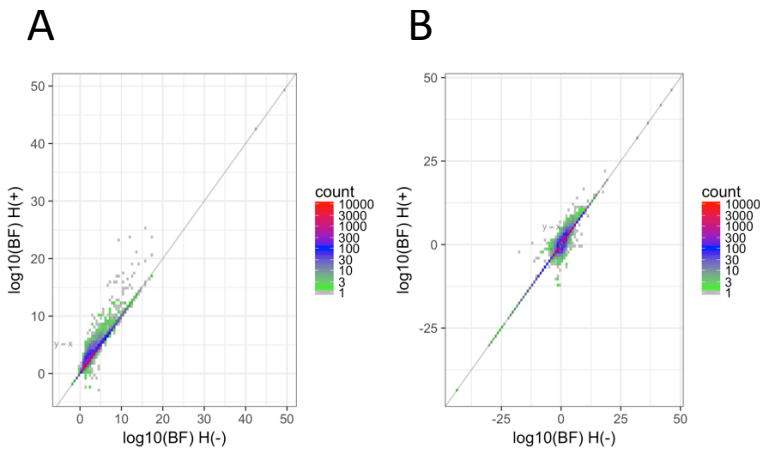

Figure 1: Horizontal axis shows the logarithm of Bayes factor when heterozygous SNPs information is not used in HapMuC generative model. Vertical axis shows the logarithm of Bayes factor when heterozygous SNPs information is used in HapMuC generative model. (**A**)Comparison of Bayes factor, when true somatic mutation exists. (**B**)Comparison of Bayes factor, when no true somatic mutation exists.

A.2 Improvement on Bayes Factor when Strand Bias Information Is Available

Using the exome sequence data of renal clear-cell carcinoma, we confirmed the sufficient condition, when strand bias information is available. OVarCall generative model is used for computing Bayes factor.

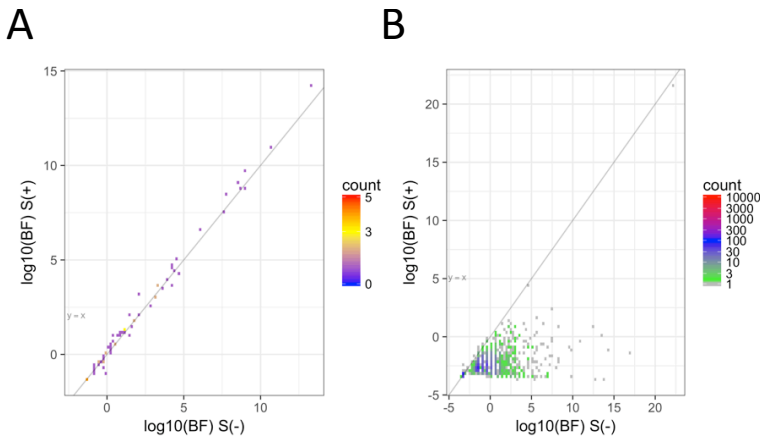

Figure 2: Horizontal axis shows the logarithm of Bayes factor when strand bias information is not used in OVarCall generative model. Vertical axis shows the logarithm of Bayes factor when strand bias information is used in OVarCall generative model. (**A**)Comparison of Bayes factor, when true somatic mutation exists. (**B**)Comparison of Bayes factor, when no true somatic mutation exists.

A.3 Improvement on Bayes Factor when Overlapping Paired-End Reads Information Is Available

Using the exome sequence data of renal clear-cell carcinoma, we confirmed the sufficient condition, when overlapping paired-end reads information is available. OVarCall generative model is used for computing Bayes factor.

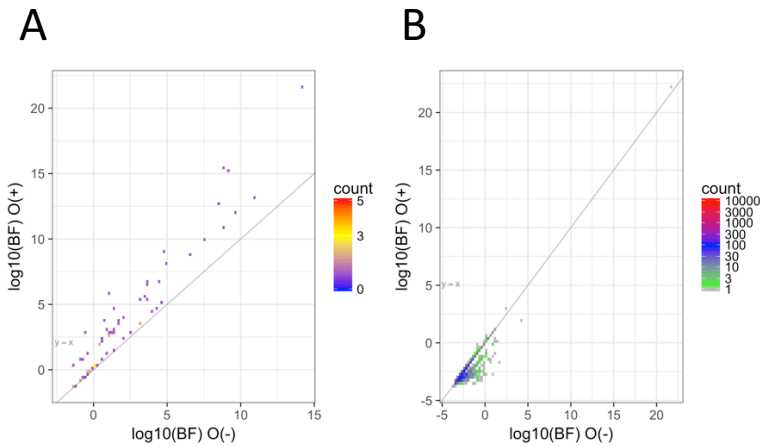

Figure 3: Horizontal axis shows the logarithm of Bayes factor when overlapping paired-end reads information is not used in OVarCall generative model. Vertical axis shows the logarithm of Bayes factor when overlapping paired-end reads information is used in OVarCall generative model. (**A**)Comparison of Bayes factor, when true somatic mutation exists. (**B**)Comparison of Bayes factor, when no true somatic mutation exists.

B Tumor Model in OHVarfinDer

B.1 Notation Summary of Tumor Model in OHVarfinDer

Table 1: Notation summary

Notation	Type	Meaning
$[N]$	set of int	{0,..,N-1}
$\boldsymbol{I}_k$	type for 1 out of k expression vector	definition of type
$Idx(\cdot)$	function for $\boldsymbol{I}_k \to$ int $\in [k]$	returns the index at which the element is one.
Δ^k	type for k dimensional non negative simplex	definition of type
$Arr(x)$	array, each element is of type x	definition of type
R_u	real number $\in [0, 1]$	definition of type
R_p	real number $\in (0, \infty)$	definition of type
n	int	order of observed paired-end reads
$\boldsymbol{R}_{+,n}$	string $\in [ATGC]^*$	n-th observed forward read
$\boldsymbol{R}_{-,n}$	string $\in [ATGC]^*$	n-th observed reverse read
$\boldsymbol{R}_{\pm,n}$	(string, string)	n-th observed paried-end read
t_n	int $\in \{0, 1, 2, 3, 4\}$	n-th type of each observed paired-end reads
$\mathcal{R}$	$Arr((string, string))$	array of paired-end reads
$\mathcal{R}_N$	$Arr((string, string))$	array of paired-end reads @ normal data
$\mathcal{R}_T$	$Arr((string, string))$	array of paired-end reads @ tumor data
i	int $\in [22]$	pattern index of latent paired-end reads
$\boldsymbol{Z}_n$	$\boldsymbol{I}_{22}$	latent variable for n-th paried-end read
$\boldsymbol{H}_{i,+}$	string $\in [ATGC]^*$	i-th latent forward strand read sequence
$\boldsymbol{H}_{i,-}$	string $\in [ATGC]^*$	i-th latent reverse strand read sequence
$\boldsymbol{H}_{i,\pm}$	(string, string)	i-th latent paired-end reads
$\mathcal{H}$	$Arr((string, string))$	all 22 patterns of latent paired-end reads
$\boldsymbol{\pi}_H$	Δ^3	parameter for haplotype frequencies with variant
π_F	R_u	parameter for reference allele frequency
ϵ_l	R_u	parameter for error rate in overlapping paired-end reads
ϵ_h	R_u	parameter for error rate in hetero SNP covering reads
ϵ_b	R_u	parameter for strand bias rate
π_{HE}	R_u	Haplotype frequency without variant
ϵ_{be}	R_u	Strand bias rate
ϵ_s	R_u	Error rate for unpaired read
$\mathcal{Z}_{tumor}$	$(Arr(\boldsymbol{I}_{22}), \Delta^3, R_u, R_u, R_u, R_u)$	definition of type
$\mathcal{Z}_{error}$	$(Arr(\boldsymbol{I}_{22}), R_u, R_u, R_u, R_u, R_u)$	definition of type
$\mathcal{Z}_T$	$\mathcal{Z}_{tumor}$	all parameters and latent variables of tumor data
$\mathcal{Z}_N$	$\mathcal{Z}_{error}$	all parameters and latent variables of normal data
$\boldsymbol{\gamma}_H$	(R_p, R_p, R_p)	hyperparameter for $\boldsymbol{\pi}_H$
$\boldsymbol{\gamma}_F$	(R_p, R_p)	hyperparameter for $\boldsymbol{\pi}_F$
$\boldsymbol{\alpha}_l$	(R_p, R_p)	hyperparameter for ϵ_l
$\boldsymbol{\alpha}_h$	(R_p, R_p)	hyperparameter for ϵ_h
$\boldsymbol{\alpha}_b$	(R_p, R_p)	hyperparameter for ϵ_b
$\boldsymbol{\gamma}_{HE}$	(R_p, R_p)	Hyperparameter for $\boldsymbol{\pi}_{HE}$
$\boldsymbol{\alpha}_{be}$	(R_p, R_p)	Hyperparameter for ϵ_{be}
$\boldsymbol{\alpha}_s$	(R_p, R_p)	Hyperparameter for ϵ_s

B.2 Meaning of the category

B.2.1 O(+)H(-) category of read pair

This category of read pair ($t_n = 0$) is overlapping between forward read and reverse read on the candidate mutation position, and covers no heterozygous SNPs nearby the candidate position.

B.2.2 O(-)H(+) category of read pair

This category of read pair ($t_n = 1$) is not overlapping between forward read and reverse read read on the candidate mutation position, and covers heterozygous SNPs nearby the candidate position.

B.2.3 O(+)H(+) category of read pair

This category of read pair ($t_n = 2$) is overlapping between forward read and reverse read read on the candidate mutation position, and covers heterozygous SNPs nearby the candidate position.

B.2.4 O(-)H(-)S(+) category of read pair

This category of read pair ($t_n = 3$) is not overlapping between forward read and reverse read read on the candidate mutation position, and covers no heterozygous SNPs nearby the candidate position. The candidate mutation position is covered by forward read. (Forward/reverse is decided by the mapping direction, compared with reference sequence.)

B.2.5 O(-)H(-)S(-) category of read pair

This category of read pair ($t_n = 4$) is not overlapping between forward read and reverse read read on the candidate mutation position, and covers no heterozygous SNPs nearby the candidate position. The candidate mutation position is covered by reverse read.

B.3 Joint Probability Density Function of Tumor Model

Here, for simplicity, we omit the hypothesis $\mathcal{H}$ and denote joint probability (and marginal likelihood) in tumor model as $P_T(\cdot)$.

$$
\begin{aligned}
&P_T(\mathcal{R}_N, \mathcal{R}_T, \mathcal{Z}_N, \mathcal{Z}_T | \boldsymbol{\gamma}_{HE}, \boldsymbol{\gamma}_H, \boldsymbol{\gamma}_F, \boldsymbol{\alpha}_s, \boldsymbol{\alpha}_l, \boldsymbol{\alpha}_h, \boldsymbol{\alpha}_b) \\
&= P_{bata}(\boldsymbol{\pi}_{HE} | \boldsymbol{\gamma}_{HE}) P_{dir}(\boldsymbol{\pi}_H | \boldsymbol{\gamma}_H) P_{beta}(\boldsymbol{\pi}_F | \boldsymbol{\gamma}_F) \\
&\quad \cdot P_{bata}(\epsilon_s | \boldsymbol{\alpha}_s) P_{bata}(\epsilon_l | \boldsymbol{\alpha}_l) P_{bata}(\epsilon_h | \boldsymbol{\alpha}_h) P_{bata}(\epsilon_b | \boldsymbol{\alpha}_b) \\
&\quad \cdot F_e(\mathcal{R}_N, \mathcal{Z}_N, \{t_{N,n}\}_n) \\
&\quad \cdot F_t(\mathcal{R}_T, \mathcal{Z}_T, \{t_{T,n}\}_n)
\end{aligned}
$$

where,

$$\mathcal{Z}_N = (\{Z_{N,n}\}_n, \pi_{HE}, \epsilon_l, \epsilon_h, \epsilon_b, \epsilon_s)$$
$$\mathcal{Z}_T = (\{Z_{T,n}\}_n, \pi_H, \pi_F, \epsilon_l, \epsilon_h, \epsilon_b)$$

$$\begin{aligned}
&\ln F_e(\mathcal{R}_N, \mathcal{Z}_N, \{t_{N,n}\}_n) \\
&= \sum_n \left[T_{(t_{N,n}),0} Z_{N,n,0} \left\{ 2\ln(1-\epsilon_l) + \ln P(\boldsymbol{R}_{N,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,0})}) \right\} \right] \\
&\quad + \sum_n \left[T_{(t_{N,n}),1} Z_{N,n,1} \left\{ 2\ln\epsilon_l + \ln P(\boldsymbol{R}_{N,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,1})}) \right\} \right] \\
&\quad + \sum_n \left[T_{(t_{N,n}),2} Z_{N,n,2} \left\{ \ln\epsilon_l(1-\epsilon_l)\epsilon_{be} + \ln P(\boldsymbol{R}_{N,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,2})}) \right\} \right] \\
&\quad + \sum_n \left[T_{(t_{N,n}),3} Z_{N,n,3} \left\{ \ln\epsilon_l(1-\epsilon_l)(1-\epsilon_{be}) + \ln P(\boldsymbol{R}_{N,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,3})}) \right\} \right] \\
&\quad + \sum_n \left[T_{(t_{N,n}),4} Z_{N,n,4} \left\{ \ln\pi_{HE,0}(1-\epsilon_h) + \ln P(\boldsymbol{R}_{N,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,4})}) \right\} \right] \\
&\quad + \sum_n \left[T_{(t_{N,n}),5} Z_{N,n,5} \left\{ \ln\pi_{HE,1}(1-\epsilon_h) + \ln P(\boldsymbol{R}_{N,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,5})}) \right\} \right] \\
&\quad + \sum_n \left[T_{(t_{N,n}),6} Z_{N,n,6} \left\{ \ln\pi_{HE,0}\epsilon_h + \ln P(\boldsymbol{R}_{N,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,6})}) \right\} \right] \\
&\quad + \sum_n \left[T_{(t_{N,n}),7} Z_{N,n,7} \left\{ \ln\pi_{HE,1}\epsilon_h + \ln P(\boldsymbol{R}_{N,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,7})}) \right\} \right] \\
&\quad + \sum_n \left[T_{(t_{N,n}),8} Z_{N,n,8} \left\{ \ln\pi_{HE,0}(1-\epsilon_l)^2 + \ln P(\boldsymbol{R}_{N,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,8})}) \right\} \right] \\
&\quad + \sum_n \left[T_{(t_{N,n}),9} Z_{N,n,9} \left\{ \ln\pi_{HE,1}(1-\epsilon_l)^2 + \ln P(\boldsymbol{R}_{N,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,9})}) \right\} \right] \\
&\quad + \sum_n \left[T_{(t_{N,n}),10} Z_{N,n,10} \left\{ \ln\pi_{HE,0}\epsilon_l^2 + \ln P(\boldsymbol{R}_{N,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,10})}) \right\} \right] \\
&\quad + \sum_n \left[T_{(t_{N,n}),11} Z_{N,n,11} \left\{ \ln\pi_{HE,1}\epsilon_l^2 + \ln P(\boldsymbol{R}_{N,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,11})}) \right\} \right] \\
&\quad + \sum_n \left[T_{(t_{N,n}),12} Z_{N,n,12} \left\{ \ln\pi_{HE,0}(1-\epsilon_l)\epsilon_l\epsilon_{be} + \ln P(\boldsymbol{R}_{N,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,12})}) \right\} \right] \\
&\quad + \sum_n \left[T_{(t_{N,n}),13} Z_{N,n,13} \left\{ \ln\pi_{HE,1}(1-\epsilon_l)\epsilon_l\epsilon_{be} + \ln P(\boldsymbol{R}_{N,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,13})}) \right\} \right] \\
&\quad + \sum_n \left[T_{(t_{N,n}),14} Z_{N,n,14} \left\{ \ln\pi_{HE,0}\epsilon_l(1-\epsilon_l)(1-\epsilon_{be}) + \ln P(\boldsymbol{R}_{N,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,14})}) \right\} \right] \\
&\quad + \sum_n \left[T_{(t_{N,n}),15} Z_{N,n,15} \left\{ \ln\pi_{HE,1}\epsilon_l(1-\epsilon_l)(1-\epsilon_{be}) + \ln P(\boldsymbol{R}_{N,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,15})}) \right\} \right]
\end{aligned}$$

$$
\begin{aligned}
&+\sum_n \left[T_{(t_{N,n}),16} Z_{N,n,16} \left\{\ln(1-\epsilon_s) + \ln P(\boldsymbol{R}_{N,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,16})})\right\}\right] \\
&+\sum_n \left[T_{(t_{N,n}),17} Z_{N,n,17} \left\{\ln \epsilon_s(1-\epsilon_{be}) + \ln P(\boldsymbol{R}_{N,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,17})})\right\}\right] \\
&+\sum_n \left[T_{(t_{N,n}),18} Z_{N,n,18} \left\{\ln \epsilon_s\epsilon_{be} + \ln P(\boldsymbol{R}_{N,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,18})})\right\}\right] \\
&+\sum_n \left[T_{(t_{N,n}),19} Z_{N,n,19} \left\{\ln(1-\epsilon_s) + \ln P(\boldsymbol{R}_{N,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,19})})\right\}\right] \\
&+\sum_n \left[T_{(t_{N,n}),20} Z_{N,n,20} \left\{\ln \epsilon_s\epsilon_{be} + \ln P(\boldsymbol{R}_{N,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,20})})\right\}\right] \\
&+\sum_n \left[T_{(t_{N,n}),21} Z_{N,n,21} \left\{\ln \epsilon_s(1-\epsilon_{be}) + \ln P(\boldsymbol{R}_{N,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{N,n,21})})\right\}\right]
\end{aligned}
$$

$$
\begin{aligned}
&\ln F_t(\mathcal{R}_T, \mathcal{Z}_T, \{t_{T,n}\}_n) \\
&=\sum_n \left[T_{(t_{T,n}),0} Z_{T,n,0} \left\{\ln \pi_F(1-\epsilon_l) + \ln P(\boldsymbol{R}_{T,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,0})})\right\}\right] \\
&+\sum_n \left[T_{(t_{T,n}),1} Z_{T,n,1} \left\{\ln(1-\pi_F) + \ln P(\boldsymbol{R}_{T,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,1})})\right\}\right] \\
&+\sum_n \left[T_{(t_{T,n}),2} Z_{T,n,2} \left\{\ln \pi_F\epsilon_l\epsilon_b + \ln P(\boldsymbol{R}_{T,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,2})})\right\}\right] \\
&+\sum_n \left[T_{(t_{T,n}),3} Z_{T,n,3} \left\{\ln \pi_F\epsilon_l(1-\epsilon_b) + \ln P(\boldsymbol{R}_{T,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,3})})\right\}\right] \\
&+\sum_n \left[T_{(t_{T,n}),4} Z_{T,n,4} \left\{\ln \pi_{H,0} + \ln P(\boldsymbol{R}_{T,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,4})})\right\}\right] \\
&+\sum_n \left[T_{(t_{T,n}),5} Z_{T,n,5} \left\{\ln \pi_{H,1}(1-\epsilon_h) + \ln P(\boldsymbol{R}_{T,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,5})})\right\}\right] \\
&+\sum_n \left[T_{(t_{T,n}),6} Z_{T,n,6} \left\{\ln \pi_{H,2} + \ln P(\boldsymbol{R}_{T,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,6})})\right\}\right] \\
&+\sum_n \left[T_{(t_{T,n}),7} Z_{T,n,7} \left\{\ln \pi_{H,1}\epsilon_h + \ln P(\boldsymbol{R}_{T,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,7})})\right\}\right] \\
&+\sum_n \left[T_{(t_{T,n}),8} Z_{T,n,8} \left\{\ln \pi_{H,0}(1-\epsilon_l) + \ln P(\boldsymbol{R}_{T,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,8})})\right\}\right] \\
&+\sum_n \left[T_{(t_{T,n}),9} Z_{T,n,9} \left\{\ln \pi_{H,1}(1-\epsilon_l)^2 + \ln P(\boldsymbol{R}_{T,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,9})})\right\}\right] \\
&+\sum_n \left[T_{(t_{T,n}),10} Z_{T,n,10} \left\{\ln \pi_{H,2} + \ln P(\boldsymbol{R}_{T,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,10})})\right\}\right] \\
&+\sum_n \left[T_{(t_{T,n}),11} Z_{T,n,11} \left\{\ln \pi_{H,1}\epsilon_l^2 + \ln P(\boldsymbol{R}_{T,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,11})})\right\}\right] \\
&+\sum_n \left[T_{(t_{T,n}),12} Z_{T,n,12} \left\{\ln \pi_{H,0}\epsilon_l\epsilon_b + \ln P(\boldsymbol{R}_{T,\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,12})})\right\}\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_n \left[T_{(t_{T,n}),13} Z_{T,n,13} \left\{ \ln \pi_{H,1}(1-\epsilon_l)\epsilon_l\epsilon_b + \ln P(\boldsymbol{R}_{T,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,13})}) \right\} \right] \\
&+ \sum_n \left[T_{(t_{T,n}),14} Z_{T,n,14} \left\{ \ln \pi_{H,0}\epsilon_l(1-\epsilon_b) + \ln P(\boldsymbol{R}_{T,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,14})}) \right\} \right] \\
&+ \sum_n \left[T_{(t_{T,n}),15} Z_{T,n,15} \left\{ \ln \pi_{H,1}\epsilon_l(1-\epsilon_l)(1-\epsilon_b) + \ln P(\boldsymbol{R}_{T,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,15})}) \right\} \right] \\
&+ \sum_n \left[T_{(t_{T,n}),16} Z_{T,n,16} \left\{ \ln \pi_F + \ln P(\boldsymbol{R}_{T,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,16})}) \right\} \right] \\
&+ \sum_n \left[T_{(t_{T,n}),17} Z_{T,n,17} \left\{ \ln(1-\pi_F)(1-\epsilon_b) + \ln P(\boldsymbol{R}_{T,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,17})}) \right\} \right] \\
&+ \sum_n \left[T_{(t_{T,n}),18} Z_{T,n,18} \left\{ \ln(1-\pi_F)\epsilon_b + \ln P(\boldsymbol{R}_{T,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,18})}) \right\} \right] \\
&+ \sum_n \left[T_{(t_{T,n}),19} Z_{T,n,19} \left\{ \ln \pi_F + \ln P(\boldsymbol{R}_{T,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,19})}) \right\} \right] \\
&+ \sum_n \left[T_{(t_{T,n}),20} Z_{T,n,20} \left\{ \ln(1-\pi_F)\epsilon_b + \ln P(\boldsymbol{R}_{T,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,20})}) \right\} \right] \\
&+ \sum_n \left[T_{(t_{T,n}),21} Z_{T,n,21} \left\{ \ln(1-\pi_F)(1-\epsilon_b) + \ln P(\boldsymbol{R}_{T,\pm,n} | \boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,21})}) \right\} \right]
\end{aligned}
$$

where, T is the 5 by 22 matrix, which has the possible patterns within 22 original latent paired-end reads given the one of 5 types of paired-end read.

$$
T_{i,j} = \begin{cases}
1 & (i = 0 \text{ (O(+)H(-))}, j \in \{0,1,2,3\}) \\
& (i = 1 \text{ (O(-)H(+))}, j \in \{4,5,6,7\}), \\
& (i = 2 \text{ (O(+)H(+))}, j \in \{8,9,10,11,12,13,14,15\}), \\
& (i = 3 \text{ (O(-)H(-)S(+))}, j \in \{16,17,18\}), \\
& (i = 4 \text{ (O(-)H(-)S(-))}, j \in \{19,20,21\}) \\
0 & \textit{otherwise}
\end{cases}
$$

B.4 Derivation of Lower Bound in Tumor Model

By applying the Jensen's inequality, lower bound for marginal likelihood is expressed as follows.

$$
\begin{aligned}
&\ln P_T(\mathcal{R}_N, \mathcal{R}_T | \boldsymbol{\gamma}_{HE}, \boldsymbol{\gamma}_H, \boldsymbol{\gamma}_F, \boldsymbol{\alpha}_s, \boldsymbol{\alpha}_l, \boldsymbol{\alpha}_h, \boldsymbol{\alpha}_b) \\
&\geq E_q \left[\ln \frac{P(\mathcal{R}, \mathcal{Z} | \boldsymbol{\gamma}_H, \boldsymbol{\gamma}_F, \boldsymbol{\alpha}_l, \boldsymbol{\alpha}_h, \boldsymbol{\alpha}_b)}{q(\{\boldsymbol{Z}_{T,n}\}_n) q(\{\boldsymbol{Z}_{N,n}\}_n) q(\pi_H) q(\pi_F) q(\epsilon_l) q(\epsilon_h) q(\epsilon_b))} \right] (= \mathcal{L}_T(q)) \\
&= E_q \left[-\ln B(\boldsymbol{\gamma}_{HE}) + (\gamma_{HE,0} - 1) \ln \pi_{HE,0} + (\gamma_{H,1} - 1) \ln(1 - \pi_{H,0}) \right] \\
&\quad + E_q \left[-\ln B(\boldsymbol{\gamma}_H) + (\gamma_{H,0} - 1) \ln \pi_{H,0} + (\gamma_{H,1} - 1) \ln \pi_{H,1} + (\gamma_{H,2} - 1) \ln \pi_{H,2} \right] \\
&\quad + E_q \left[-\ln B(\boldsymbol{\gamma}_F) + (\gamma_{F,0} - 1) \ln \pi_{F,0} + (\gamma_{F,1} - 1) \ln(1 - \pi_F) \right] \\
&\quad + E_q \left[-\ln B(\boldsymbol{\alpha}_s) + (\alpha_{s,0} - 1) \ln \epsilon_s + (\alpha_{s,1} - 1) \ln(1 - \epsilon_s) \right]
\end{aligned}
$$

$$
\begin{aligned}
&+ E_q\left[-\ln B(\boldsymbol{\alpha}_l) + (\alpha_{l,0}-1)\ln\epsilon_l + (\alpha_{l,1}-1)\ln(1-\epsilon_l)\right] \\
&+ E_q\left[-\ln B(\boldsymbol{\alpha}_h) + (\alpha_{h,0}-1)\ln\epsilon_h + (\alpha_{h,1}-1)\ln(1-\epsilon_h)\right] \\
&+ E_q\left[-\ln B(\boldsymbol{\alpha}_b) + (\alpha_{b,0}-1)\ln\epsilon_b + (\alpha_{b,1}-1)\ln(1-\epsilon_b)\right] \\
&+ E_q\left[\ln F_e(\mathcal{R}_N, \mathcal{Z}_N, \{t_{N,n}\}_n)\right] \\
&+ E_q\left[\ln F_t(\mathcal{R}_T, \mathcal{Z}_T, \{t_{T,n}\}_n)\right] \\
&- E_q\left[\ln(q(\{\boldsymbol{Z}_{T,n}\}_n)q(\{\boldsymbol{Z}_{N,n}\}_n)q(\pi_H)q(\pi_F)q(\epsilon_l)q(\epsilon_h)q(\epsilon_b))\right]
\end{aligned}
$$

,where $q(\pi_H)q(\pi_F)q(\epsilon_l)q(\epsilon_h)q(\epsilon_b)q(\{\boldsymbol{Z}_{T,n}\}_n)q(\{\boldsymbol{Z}_{N,n}\}_n)$ represents any probability density function for each parameters or latent variables. As in the usual variational Bayes procedures, here we will maximize the lower bound $\mathcal{L}_T(q)$ with respect to each variational distributions.

B.4.1 Maximize $\mathcal{L}_T(q)$ with respect to $q(\boldsymbol{\pi}_H)$

Following the variational Bayes procedure, we would like to get optimal $q^*(\boldsymbol{\pi}_H)$ which maximize $\mathcal{L}_T(q)$ with respect to $q(\boldsymbol{\pi}_H)$.

$$
\begin{aligned}
\mathcal{L}_T(q) &= E_q\left[\left\{(\gamma_{H,0}-1) + \sum_n \sum_{j\in\{4,8,12,14\}} T_{t_{T,n},j} Z_{T,n,j}\right\}\ln\pi_{H,0}\right] \\
&\quad + E_q\left[\left\{(\gamma_{H,1}-1) + \sum_n \sum_{j\in\{5,7,9,11,13,15\}} T_{t_{T,n},j} Z_{T,n,j}\right\}\ln\pi_{H,1}\right] \\
&\quad + E_q\left[\left\{(\gamma_{H,2}-1) + \sum_n \sum_{j\in\{6,10\}} T_{t_{T,n},j} Z_{T,n,j}\right\}\ln\pi_{H,2}\right] \\
&\quad - E_q\left[\ln q(\boldsymbol{\pi}_H)\right] + Const. \\
&= -KL[q(\boldsymbol{\pi}_H)||p_{dir}(\boldsymbol{\pi}_H|\boldsymbol{\gamma}_H^*)] + Const.
\end{aligned}
$$

where,

$$
\begin{aligned}
\gamma_{H,0}^* &= E_q\left[(\gamma_{H,0}-1) + \sum_n \sum_{j\in\{4,8,12,14\}} T_{t_{T,n},j} Z_{T,n,j}\right] \\
\gamma_{H,1}^* &= E_q\left[(\gamma_{H,1}-1) + \sum_n \sum_{j\in\{5,7,9,11,13,15\}} T_{t_{T,n},j} Z_{T,n,j}\right] \\
\gamma_{H,2}^* &= E_q\left[(\gamma_{H,2}-1) + \sum_n \sum_{j\in\{6,10\}} T_{t_{T,n},j} Z_{T,n,j}\right]
\end{aligned}
$$

Therefore, we can maximize the lower bound by minimization of KL divergence of $KL[q(\boldsymbol{\pi}_H)||p_{dir}(\boldsymbol{\pi_H}|\boldsymbol{\gamma}_H^*)] \geq 0$. The optimal form distribution

is

$$q^*(\boldsymbol{\pi}_H) = p_{dir}(\boldsymbol{\pi_H}|\boldsymbol{\gamma}_H^*)$$

B.4.2 Maximize $\mathcal{L}_T(q)$ with respect to $q(\boldsymbol{\pi}_F)$

We would like to get optimal $q^*(\pi_F)$ which maximize $\mathcal{L}_T(q)$ with respect to $q(\pi_F)$.

$$\begin{aligned}
\mathcal{L}_T(q) &= E_q\left[\left\{(\gamma_{F,0}-1) + \sum_n \sum_{j\in\{0,2,3,16,19\}} T_{t_{T,n},j} Z_{T,n,j}\right\} \ln \pi_F\right] \\
&\quad + E_q\left[\left\{(\gamma_{F,1}-1) + \sum_n \sum_{j\in\{1,17,18,20,21\}} T_{t_{T,i,n},j} Z_{T,n,j}\right\} \ln(1-\pi_F)\right] \\
&\quad - E_q\left[\ln q(\boldsymbol{\pi}_F)\right] + Const. \qquad (3) \\
&= -KL[q(\pi_F)||p_{beta}(\pi_F|\boldsymbol{\gamma}_F^*)] + Const.
\end{aligned}$$

where,

$$\begin{aligned}
\gamma_{F,0}^* &= E_q\left[(\gamma_{F,0}-1) + \sum_n \sum_{j\in\{0,2,3,16,19\}} T_{t_{T,n},j} Z_{T,n,j}\right] \\
\gamma_{F,1}^* &= E_q\left[(\gamma_{F,1}-1) + \sum_n \sum_{j\in\{1,17,18,20,21\}} T_{t_{T,n},j} Z_{T,n,j}\right]
\end{aligned}$$

Therefore, we can maximize the lower bound by minimization of KL divergence of $KL[q(\pi_H)||p_{beta}(\pi_F|\boldsymbol{\gamma}_F^*)] \geq 0$. The optimal form distribution is

$$q^*(\pi_F) = p_{beta}(\pi_F|\boldsymbol{\gamma}_F^*)$$

B.4.3 Maximize $\mathcal{L}_T(q)$ with respect to $q(\epsilon_l)$

We would like to get optimal $q^*(\epsilon_l)$ which maximize $\mathcal{L}_T(q)$ with respect to $q(\epsilon_l)$.

$$
\begin{aligned}
\mathcal{L}_T(q) &= E_q\left[\{(\alpha_{l,0}-1)\}\ln\epsilon_l\right] \\
&\quad + E_q\left[\left\{\sum_n \sum_{j\in\{2,3,11,12,13,14\}} T_{t_{T,n},j} Z_{T,n,j}\right\}\ln\epsilon_l\right] \\
&\quad + E_q\left[\left\{\sum_n \sum_{j\in\{4,6,8,19,12,14\}} T_{t_{N,n},j} Z_{N,n,j}\right\}\ln\epsilon_l\right] \\
&\quad + E_q\left[\{(\alpha_{l,1}-1)\}\ln(1-\epsilon_l)\right] \\
&\quad + E_q\left[\left\{\sum_n \sum_{j\in\{0,8,9,15\}} T_{t_{T,n},j} Z_{T,n,j}\right\}\ln(1-\epsilon_l)\right] \\
&\quad + E_q\left[\left\{\sum_n \sum_{j\in\{5,7,9,11,13,15\}} T_{t_{N,n},j} Z_{N,n,j}\right\}\ln(1-\epsilon_l)\right] \\
&\quad - E_q\left[\ln q(\boldsymbol{\epsilon}_l)\right] + Const. \\
&= -KL[q(\epsilon_l)||p_{beta}(\epsilon_l|\boldsymbol{\alpha}_l^*)] + Const.
\end{aligned}
$$

where,

$$
\begin{aligned}
\alpha_{l,0}^* &= E_q\left[(\alpha_{l,0}-1) + \sum_n \sum_{j\in\{2,3,11,12,13,14\}} T_{t_{T,n},j} Z_{T,n,j}\right] \\
&\quad + E_q\left[\sum_n \sum_{j\in\{4,6,8,19,12,14\}} T_{t_{N,n},j} Z_{N,n,j}\right] \\
\alpha_{l,1}^* &= E_q\left[(\alpha_{l,1}-1) + \sum_n \sum_{j\in\{0,8,9,15\}} T_{t_{T,n},j} Z_{T,n,j}\right] \\
&\quad + E_q\left[\sum_n \sum_{j\in\{5,7,9,11,13,15\}} T_{t_{N,n},j} Z_{N,n,j}\right]
\end{aligned}
$$

Therefore, we can maximize the lower bound by minimization of KL divergence of $KL[q(\epsilon_l)||p_{beta}(\epsilon_l|\boldsymbol{\alpha}_l^*)] \geq 0$. The optimal form distribution is

$$
q^*(\epsilon_l) = p_{beta}(\epsilon_l|\boldsymbol{\alpha}_l^*)
$$

B.4.4 Maximize $\mathcal{L}_T(q)$ with respect to $q(\epsilon_h)$

We would like to get optimal $q^*(\epsilon_h)$ which maximize $\mathcal{L}_T(q)$ with respect to $q(\epsilon_h)$.

$$
\begin{aligned}
\mathcal{L}_T(q) = & E_q\left[\left\{(\alpha_{l,0}-1)+\sum_n \sum_{j\in\{7,11,15\}} T_{t_{T,n},j} Z_{T,n,j}\right\}\ln\epsilon_l\right] \\
& + E_q\left[\left\{\sum_n \sum_{j\in\{6,7,10,11,14,15\}} T_{t_{N,n},j} Z_{N,n,j}\right\}\ln\epsilon_l\right] \\
& + E_q\left[\left\{(\alpha_{l,1}-1)+\sum_n \sum_{j\in\{5,9,13\}} T_{t_{T,n},j} Z_{T,n,j}\right\}\ln(1-\epsilon_l)\right] \\
& + E_q\left[\left\{\sum_n \sum_{j\in\{4,5,8,9,12,13\}} T_{t_{N,n},j} Z_{N,n,j}\right\}\ln(1-\epsilon_l)\right] \\
& - E_q\left[\ln q(\boldsymbol{\epsilon}_h)\right] + Const. \\
= & -KL[q(\epsilon_l)||p_{beta}(\epsilon_l|\boldsymbol{\alpha}_l^*)] + Const.
\end{aligned}
$$

where,

$$
\begin{aligned}
\alpha_{h,0}^* = & E_q\left[(\alpha_{h,0}-1)+\sum_n \sum_{j\in\{7,11,15\}} T_{t_{T,i,n},j} Z_{T,n,j}\right] \\
& + E_q\left[\sum_n \sum_{j\in\{6,7,10,11,14,15\}} T_{t_{N,n},j} Z_{N,n,j}\right] \\
\alpha_{h,1}^* = & E_q\left[(\alpha_{h,1}-1)+\sum_n \sum_{j\in\{5,9,13\}} T_{t_{T,n},j} Z_{T,n,j}\right] \\
& + E_q\left[\sum_n \sum_{j\in\{4,5,8,9,12,13\}} T_{t_{N,n},j} Z_{N,n,j}\right]
\end{aligned}
\tag{4}
$$

Therefore, we can maximize the lower bound by minimization of KL divergence of $KL[q(\epsilon_h)||p_{beta}(\epsilon_h|\boldsymbol{\alpha}_h^*)]$. The optimal form distribution is

$$q^*(\epsilon_h) = p_{beta}(\epsilon_h|\boldsymbol{\alpha}_h^*)$$

B.4.5 Maximize $\mathcal{L}_T(q)$ with respect to $q(\epsilon_b)$

We would like to get optimal $q^*(\epsilon_b)$ which maximize $\mathcal{L}_T(q)$ with respect to $q(\epsilon_b)$.

$$
\begin{aligned}
\mathcal{L}_T(q) = & E_q\left[\left\{(\alpha_{b,0}-1)+\sum_n \sum_{j\in\{2,12,13,18,20\}} T_{t_{N,n},j} Z_{N,n,j}\right\}\ln\epsilon_b\right] \\
& + E_q\left[\left\{(\alpha_{b,1}-1)+\sum_n \sum_{j\in\{3,14,15,17,21\}} T_{t_{N,n},j} Z_{N,n,j}\right\}\ln(1-\epsilon_b)\right] \\
& - E_q\left[\ln q(\boldsymbol{\epsilon}_b)\right] + Const. \\
= & -KL[q(\epsilon_b)||p_{beta}(\epsilon_b|\boldsymbol{\alpha}_b^*)] + Const.
\end{aligned}
$$

where,

$$
\begin{aligned}
\alpha_{b,0}^* &= E_q\left[(\alpha_{b,0}-1)+\sum_n \sum_{j\in\{2,12,13,18,20\}} T_{t_{N,n},j} Z_{N,n,j}\right] \\
\alpha_{b,1}^* &= E_q\left[(\alpha_{b,1}-1)+\sum_n \sum_{j\in\{3,14,15,17,21\}} T_{t_{N,n},j} Z_{N,n,j}\right]
\end{aligned}
$$

Therefore, we can maximize the lower bound by minimization of KL divergence of $KL[q(\epsilon_b)||p_{beta}(\epsilon_b|\boldsymbol{\alpha}_b^*)] \geq 0$. The optimal form distribution is

$$
q^*(\epsilon_b) = p_{beta}(\epsilon_b|\boldsymbol{\alpha}_b^*)
$$

B.4.6 Maximize $\mathcal{L}_T(q)$ with respect to $q(\boldsymbol{Z}_{T,n})$

We would like to get optimal $q^*(\boldsymbol{Z}_{T,n})$ which maximize $\mathcal{L}_T(q)$ with respect to $q(\boldsymbol{Z}_{T,n})$.

$$
\begin{aligned}
\mathcal{L}_T(q) = & \sum_n E_q\left[Z_{T,n,0} T_{t_n,0}\left\{\ln\pi_F(1-\epsilon_l) + \ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,0})})\right\}\right] \\
& + \sum_n E_q\left[Z_{T,n,1} T_{t_n,1}\left\{\ln(1-\pi_F) + \ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,1})})\right\}\right] \\
& + \sum_n E_q\left[Z_{T,n,2} T_{t_n,2}\left\{\ln\pi_F\epsilon_l\epsilon_b + \ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,2})})\right\}\right] \\
& + \sum_n E_q\left[Z_{T,n,3} T_{t_n,3}\left\{\ln\pi_F\epsilon_l(1-\epsilon_b) + \ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,3})})\right\}\right] \\
& + \sum_n E_q\left[Z_{T,n,4} T_{t_n,4}\left\{\ln\pi_{H,0} + \ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,4})})\right\}\right] \\
& + \sum_n E_q\left[Z_{T,n,5} T_{t_n,5}\left\{\ln\pi_{H,1}(1-\epsilon_h) + \ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,5})})\right\}\right] \\
& + \sum_n E_q\left[Z_{T,n,6} T_{t_n,6}\left\{\ln\pi_{H,2} + \ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,6})})\right\}\right]
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_n E_q\left[Z_{T,n,7}T_{t_n,7}\left\{\ln\pi_{H,1}\epsilon_h+\ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,7})})\right\}\right]\\
&+\sum_n E_q\left[Z_{T,n,8}T_{t_n,8}\left\{\ln\pi_{H,0}(1-\epsilon_l)+\ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,8})})\right\}\right]\\
&+\sum_n E_q\left[Z_{T,n,9}T_{t_n,9}\left\{\ln\pi_{H,1}(1-\epsilon_l)^2+\ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,9})})\right\}\right]\\
&+\sum_n E_q\left[Z_{T,n,10}T_{t_n,10}\left\{\ln\pi_{H,2}+\ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,10})})\right\}\right]\\
&+\sum_n E_q\left[Z_{T,n,11}T_{t_n,11}\left\{\ln\pi_{H,1}\epsilon_l^2+\ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,11})})\right\}\right]\\
&+\sum_n E_q\left[Z_{T,n,12}T_{t_n,12}\left\{\ln\pi_{H,0}\epsilon_l\epsilon_b+\ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,12})})\right\}\right]\\
&+\sum_n E_q\left[Z_{T,n,13}T_{t_n,13}\left\{\ln\pi_{H,1}(1-\epsilon_l)\epsilon_l\epsilon_b+\ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,13})})\right\}\right]\\
&+\sum_n E_q\left[Z_{T,n,14}T_{t_n,14}\left\{\ln\pi_{H,0}\epsilon_l(1-\epsilon_b)+\ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,14})})\right\}\right]\\
&+\sum_n E_q\left[Z_{T,n,15}T_{t_n,15}\left\{\ln\pi_{H,1}\epsilon_l(1-\epsilon_l)(1-\epsilon_b)+\ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,15})})\right\}\right]\\
&+\sum_n E_q\left[Z_{T,n,16}T_{t_n,16}\left\{\ln\pi_F+\ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,16})})\right\}\right]\\
&+\sum_n E_q\left[Z_{T,n,17}T_{t_n,17}\left\{\ln(1-\pi_F)(1-\epsilon_b)+\ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,17})})\right\}\right]\\
&+\sum_n E_q\left[Z_{T,n,18}T_{t_n,18}\left\{\ln(1-\pi_F)\epsilon_b+\ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,18})})\right\}\right]\\
&+\sum_n E_q\left[Z_{T,n,19}T_{t_n,19}\left\{\ln\pi_F+\ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,19})})\right\}\right]\\
&+\sum_n E_q\left[Z_{T,n,20}T_{t_n,20}\left\{\ln(1-\pi_F)\epsilon_b+\ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,20})})\right\}\right]\\
&+\sum_n E_q\left[Z_{T,n,21}T_{t_n,21}\left\{\ln(1-\pi_F)(1-\epsilon_b)+\ln P(\boldsymbol{R}_{\pm,n}|\boldsymbol{H}_{Idx(\boldsymbol{Z}_{T,n,21})})\right\}\right]\\
&-E_q\left[\ln q(\boldsymbol{Z}_{T,n})\right]+Const.\\
=&\sum_n E_q\left[Z_{T,n,0}T_{t_n,0}\ln\rho^*_{T,n,0}\right]+\sum_n E_q\left[Z_{T,n,1}T_{t_n,1}\ln\rho^*_{T,n,1}\right]\\
&+\sum_n E_q\left[Z_{T,n,2}T_{t_n,2}\ln\rho^*_{T,n,2}\right]+\sum_n E_q\left[Z_{T,n,3}T_{t_n,3}\ln\rho^*_{T,n,3}\right]\\
&+\sum_n E_q\left[Z_{T,n,4}T_{t_n,4}\ln\rho^*_{T,n,4}\right]+\sum_n E_q\left[Z_{T,n,5}T_{t_n,5}\ln\rho^*_{T,n,5}\right]\\
&+\sum_n E_q\left[Z_{T,n,6}T_{t_n,6}\ln\rho^*_{T,n,6}\right]+\sum_n E_q\left[Z_{T,n,7}T_{t_n,7}\ln\rho^*_{T,n,7}\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_n E_q\left[Z_{T,n,8} T_{t_n,8} \ln \rho^*_{T,n,8}\right] + \sum_n E_q\left[Z_{T,n,9} T_{t_n,9} \ln \rho^*_{T,n,9}\right] \\
&+ \sum_n E_q\left[Z_{T,n,10} T_{t_n,10} \ln \rho^*_{T,n,10}\right] + \sum_n E_q\left[Z_{T,n,11} T_{t_n,11} \ln \rho^*_{T,n,11}\right] \\
&+ \sum_n E_q\left[Z_{T,n,12} T_{t_n,12} \ln \rho^*_{T,n,12}\right] + \sum_n E_q\left[Z_{T,n,13} T_{t_n,13} \ln \rho^*_{T,n,13}\right] \\
&+ \sum_n E_q\left[Z_{T,n,14} T_{t_n,14} \ln \rho^*_{T,n,14}\right] + \sum_n E_q\left[Z_{T,n,15} T_{t_n,15} \ln \rho^*_{T,n,15}\right] \\
&+ \sum_n E_q\left[Z_{T,n,16} T_{t_n,16} \ln \rho^*_{T,n,16}\right] + \sum_n E_q\left[Z_{T,n,17} T_{t_n,17} \ln \rho^*_{T,n,17}\right] \\
&+ \sum_n E_q\left[Z_{T,n,18} T_{t_n,18} \ln \rho^*_{T,n,18}\right] + \sum_n E_q\left[Z_{T,n,19} T_{t_n,19} \ln \rho^*_{T,n,19}\right] \\
&+ \sum_n E_q\left[Z_{T,n,20} T_{t_n,20} \ln \rho^*_{T,n,20}\right] + \sum_n E_q\left[Z_{T,n,21} T_{t_n,21} \ln \rho^*_{T,n,21}\right] \\
&- E_q\left[\ln q(\boldsymbol{Z}_{T,n})\right] + Const. \\
&= -KL[q(\boldsymbol{Z}_{T,n}) || p_{multi}(\boldsymbol{Z}_{T,n} | \boldsymbol{\zeta}^*_{T,n})] + Const.
\end{aligned}
$$

where,

$$
\begin{aligned}
&\zeta^*_{T,n,j} \propto \rho^*_{T,n,j} \quad (when \;\; T_{t_n,j} = 1) \\
&\zeta^*_{T,n,j} = 0 \qquad\quad (when \;\; T_{t_n,j} = 0) \\
&\sum_{j=0}^{21} \zeta^*_{T,n,j} = 1
\end{aligned}
$$

Therefore, we can maximize the lower bound by minimization of KL divergence of $KL[q(\boldsymbol{Z}_{T,n}) || p_{multi}(\boldsymbol{Z}_{T,n} | \boldsymbol{\zeta}^*_{T,n})] \geq 0$. The optimal form distribution is

$$
q^*(\boldsymbol{Z}_{T,n}) = p_{multi}(\boldsymbol{Z}_{T,n} | \boldsymbol{\zeta}^*_{T,n})
$$

B.4.7 Maximize $\mathcal{L}_T(q)$ with respect to $q(\boldsymbol{Z}_{N,n})$

We would like to get optimal $q^*(\boldsymbol{Z}_{N,n})$ which maximize $\mathcal{L}_T(q)$ with respect to $q(\boldsymbol{Z}_{N,n})$. In this part, the update procedure is same as the procedure in the $q(\boldsymbol{Z}_{T,n})$, except for the parameter settings.

C Error Model in OHVarfinDer

As for the error model, we prepared two types of model. First model assumes that error rate is dependent between tumor and normal sequence data, and second model assumes that error rate is independent beween tumor and normal sequence data. Depending on the depth coverage around the candidate

mutation position, we changed the error model. We used first model when depth coverage is ≤ 100, and we used the second model otherwise. The reason for constructing two model is that the proportion of moderate low base quality reads (≥ 10, < 15) changes depending on the depth coverage. In lower depth condition, proportion of such moderate low base quality reads is relatively high both in tumor and normal sequence data. In higher depth condition, proportion of such such moderate low base quality reads is relatively high in tumor sequence data but relatively low in normal sequence data. (If the proportion of the moderate low base quality reads are high and depth is high in normal sequence data, this candidate position tends to be filtered by the number of variant reads condition in normal sequence data.) This difference induces two models of error dependent model and error independent model.

C.1 Notation Summary of Error Models in OHVarfinDer

Table 2: Notation summary

Notation	Type	Meaning
$\mathcal{Z}_{TI}$	$\mathcal{Z}_{error}$	all parameters and latent variables of tumor data @ independent error model
$\mathcal{Z}_{TD}$	$\mathcal{Z}_{error}$	all parameters and latent variables of tumor data @ dependent error model
$\mathcal{Z}_{NI}$	$\mathcal{Z}_{error}$	all parameters and latent variables of normal data @ independent error model
$\mathcal{Z}_{ND}$	$\mathcal{Z}_{error}$	all parameters and latent variables of normal data @ dependent error model
π_{HEE}	R_u	Haplotype frequency without variant
ϵ_{le}	real number R_u	Overlapping read pair specific error rate
ϵ_{he}	real number R_u	Haplotype specific error rate
ϵ_{be}	real number R_u	Strand bias rate
ϵ_{se}	R_u	Error rate for unpaired read
$\boldsymbol{\gamma}_{HEE}$	(R_p, R_p)	Hyperparameter for $\boldsymbol{\pi}_{HEE}$
$\boldsymbol{\alpha}_{se}$	real vector (R_p, R_p)	Hyperparameter for ϵ_{se}
$\boldsymbol{\alpha}_{le}$	real vector (R_p, R_p)	Hyperparameter for ϵ_{le}
$\boldsymbol{\alpha}_{he}$	real vector (R_p, R_p)	Hyperparameter for ϵ_{he}
$\boldsymbol{\alpha}_{be}$	real vector (R_p, R_p)	Hyperparameter for ϵ_{be}
$\boldsymbol{\alpha}_{se}$	real vector (R_p, R_p)	Hyperparameter for ϵ_{se}

C.2 Joint Probability Density Function of Dependent Error Model

Here, for simplicity, we omit the hypothesis $\mathcal{H}$ and denote joint probability (and marginal likelihood) in error model as $P_E(\cdot)$.

$$
\begin{aligned}
&P_E(\mathcal{R}_N, \mathcal{R}_T, \mathcal{Z}_{ND}, \mathcal{Z}_{TD} | \boldsymbol{\gamma}_{HE}, \boldsymbol{\alpha}_s, \boldsymbol{\alpha}_l, \boldsymbol{\alpha}_h, \boldsymbol{\alpha}_b) \\
&= P_{bata}(\boldsymbol{\pi}_{HE} | \boldsymbol{\gamma}_{HE}) \\
&\quad \cdot P_{bata}(\epsilon_s | \boldsymbol{\alpha}_s) P_{bata}(\epsilon_l | \boldsymbol{\alpha}_l) P_{bata}(\epsilon_h | \boldsymbol{\alpha}_h) P_{bata}(\epsilon_b | \boldsymbol{\alpha}_b) \\
&\quad \cdot F_e(\mathcal{R}_N, \mathcal{Z}_{ND}, \{\boldsymbol{t}_{N,n}\}_n) \\
&\quad \cdot F_e(\mathcal{R}_T, \mathcal{Z}_{TD}, \{\boldsymbol{t}_{T,n}\}_n)
\end{aligned}
$$

where,

$$\mathcal{Z}_{ND} = (\{Z_{N,n}\}_n, \pi_{HE}, \epsilon_l, \epsilon_h, \epsilon_s, \epsilon_b)$$
$$\mathcal{Z}_{TD} = (\{Z_{T,n}\}_n, \pi_{HE}, \epsilon_l, \epsilon_h, \epsilon_s, \epsilon_b)$$

C.3 Joint Probability Density Function of Independent Error Model

$$\begin{aligned}
&P_E(\mathcal{R}_N, \mathcal{R}_T, \mathcal{Z}_{NI}, \mathcal{Z}_{TI} | \boldsymbol{\gamma}_{HE}, \boldsymbol{\alpha}_s, \boldsymbol{\alpha}_l, \boldsymbol{\alpha}_h, \boldsymbol{\alpha}_b) \\
&= P_{bata}(\boldsymbol{\pi}_{HE} | \boldsymbol{\gamma}_{HE}) \\
&\quad \cdot P_{bata}(\epsilon_s | \boldsymbol{\alpha}_s) P_{bata}(\epsilon_l | \boldsymbol{\alpha}_l) P_{bata}(\epsilon_h | \boldsymbol{\alpha}_h) P_{bata}(\epsilon_b | \boldsymbol{\alpha}_b) \\
&\quad \cdot P_{bata}(\boldsymbol{\pi}_{HEE} | \boldsymbol{\gamma}_{HEE}) \\
&\quad \cdot P_{bata}(\epsilon_{se} | \boldsymbol{\alpha}_{se}) P_{bata}(\epsilon_{le} | \boldsymbol{\alpha}_{le}) P_{bata}(\epsilon_{he} | \boldsymbol{\alpha}_{he}) P_{bata}(\epsilon_{be} | \boldsymbol{\alpha}_{be}) \\
&\quad \cdot F_e(\mathcal{R}_N, \mathcal{Z}_{NI}, \{\boldsymbol{t}_{N,n}\}_n) \\
&\quad \cdot F_e(\mathcal{R}_T, \mathcal{Z}_{TI}, \{\boldsymbol{t}_{T,n}\}_n)
\end{aligned}$$

where,

$$\mathcal{Z}_{NI} = (\{Z_{N,n}\}_n, \pi_{HEE}, \epsilon_{le}, \epsilon_{he}, \epsilon_{se}, \epsilon_{be})$$
$$\mathcal{Z}_{TI} = (\{Z_{T,n}\}_n, \pi_{HE}, \epsilon_l, \epsilon_h, \epsilon_s, \epsilon_b)$$

C.4 Variational Bayes Procedure for Error Model

The variational bayes procedure for error model can be done in the same way as in the tumor model.

C.5 Prior Hyperparameters Used in Performance Evaluation

Table 3: Prior hyperparameters summary

Experiment	Model	Hyperparameters	Depth < 100	Depth ≥ 100
Simulation	Error model	$\boldsymbol{\gamma}_{HE}$	$(5.0, 5.0)$	"
		$\boldsymbol{\alpha}_l$	$(2.0, 30.0)$	"
		$\boldsymbol{\alpha}_h$	$(2.0, 30.0)$	"
		$\boldsymbol{\alpha}_b$	$(0.5, 0.5)$	$(0.05, 0.05)$
		$\boldsymbol{\alpha}_s$	$(2.0, 30.0)$	"
		$\boldsymbol{\gamma}_{HEE}$	$(5.0, 5.0)$	$(50.0, 50.0)$
		$\boldsymbol{\alpha}_{le}$	$(2.0, 30.0)$	"
		$\boldsymbol{\alpha}_{he}$	$(2.0, 30.0)$	"
		$\boldsymbol{\alpha}_{be}$	$(0.5, 0.5)$	$(0.05, 0.05)$
		$\boldsymbol{\alpha}_{se}$	$(2.0, 30.0)$	"
	Tumor model	$\boldsymbol{\gamma}_H$	$(5.0, 5.0, 1.0)$	"
		$\boldsymbol{\gamma}_F$	$(10.0, 1.0)$	"
		$\boldsymbol{\alpha}_l$	$(1.0, 100.0)$	"
		$\boldsymbol{\alpha}_h$	$(1.0, 100.0)$	"
		$\boldsymbol{\alpha}_b$	$(5.0, 5.0)$	$(50.0, 50.0)$
Real data	Error model	$\boldsymbol{\gamma}_{HE}$	$(2.5, 2.5)$	"
		$\boldsymbol{\alpha}_l$	$(1.0, 10.0)$	"
		$\boldsymbol{\alpha}_h$	$(1.0, 10.0)$	"
		$\boldsymbol{\alpha}_b$	$(1.0, 1.0)$	$(10.0, 10.0)$
		$\boldsymbol{\alpha}_s$	$(1.0, 10.0)$	"
		$\boldsymbol{\gamma}_{HEE}$	$(2.5, 2.5)$	"
		$\boldsymbol{\alpha}_{le}$	$(1.0, 10.0)$	"
		$\boldsymbol{\alpha}_{he}$	$(1.0, 10.0)$	"
		$\boldsymbol{\alpha}_{be}$	$(1.0, 1.0)$	$(10.0, 10.0)$
		$\boldsymbol{\alpha}_{se}$	$(1.0, 10.0)$	"
	Tumor model	$\boldsymbol{\gamma}_H$	$(2.5, 2.5, 1.0)$	"
		$\boldsymbol{\gamma}_F$	$(5.0, 1.0)$	"
		$\boldsymbol{\alpha}_l$	$(0.1, 10.0)$	"
		$\boldsymbol{\alpha}_h$	$(0.1, 10.0)$	"
		$\boldsymbol{\alpha}_b$	$(0.5, 0.5)$	$(0.05, 0.05)$

D Details of result

D.1 Filter Conditions for Simulation Data Sets

We retained the candidate positions if they met with the following criteria.

1. The read coverage is ≥ 12
2. The non-reference allele frequency in tumor sample is ≥ 0.05
3. The normal allele frequency in normal sample is ≤ 0.1
4. Variant supporting read in tumor sample is ≥ 4.

D.2 Filter Conditions for Exome Sequence Data

We retained the candidate positions for lower variant allele frequency mutations if they met with the following criteria.

1. The read coverage is ≥ 100
2. The non-reference allele frequency in tumor sample is ≥ 0.07.
3. The normal allele frequency in normal sample is ≤ 0.02
4. Variant supporting read in tumor sample is ≥ 3.
5. Variant supporting read in normal sample is ≤ 1.
6. Tri allelic frequency in tumor sample is ≤ 0.03.
7. Tri allelic read in tumor sample is ≤ 2.
8. Average Base quality in tumor samples is ≥ 25.
9. Average Base quality in normal samples is ≥ 25.
10. Distance of nearest InDel is > 25.

We retained the candidate positions for moderate variant allele frequency mutations if they met with the following criteria.

1. The read coverage is ≥ 30
2. The non-reference allele frequency in tumor sample is ≥ 0.07.
3. The normal allele frequency in normal sample is ≤ 0.02
4. Variant supporting read in tumor sample is ≥ 3.
5. Variant supporting read in normal sample is ≤ 1.
6. Tri allelic frequency in tumor sample is ≤ 0.03.
7. Tri allelic read in tumor sample is ≤ 2.
8. Average Base quality in tumor samples is ≥ 25.
9. Average Base quality in normal samples is ≥ 25.
10. Distance of nearest InDel is > 25.

We should also note that potential mapping errors are excluded by using genomic super duplications, simple repeats, and dbSNP138.

1. genomic super duplications: `http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/simpleRepeat.txt.gz`

2. simple repeats: `http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/genomicSuperDups.txt.gz`
3. dbSNP138: `http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/snp138.txt.gz`

D.3 Filter Conditions for Whole Genome Sequence Data

We retained the candidate positions if they met with the following criteria.

1. The read coverage is ≥ 10
2. The non-reference allele frequency in tumor sample is ≥ 0.05.
3. The normal allele frequency in normal sample is ≤ 0.1
4. Variant supporting read in tumor sample is ≥ 4.
5. Variant supporting read in normal sample is ≤ 1.
6. Average Base quality in tumor samples is ≥ 25.
7. Average Base quality in normal samples is ≥ 25.
8. Tri allelic frequency in tumor sample is ≤ 0.03.
9. Tri allelic read in tumor sample is ≤ 2.
10. Distance of nearest InDel is > 25.

We should also note that potential mapping errors are excluded by using genomic super duplications, simple repeats, and dbSNP138.

D.4 The Criteria for Calling Somatic Mutation on the Pure Datasets of the TCGA Mutation Calling Benchmark 4 Datasets

1. The read coverage is ≥ 15
2. The non-reference allele frequency in tumor sample is ≥ 0.15.
3. Variant supporting read in tumor sample is ≥ 6.
4. Variant supporting read in normal sample is ≤ 1.
5. Average Mapping quality in both samples is ≥ 15.
6. Average Base quality in tumor samples is ≥ 25.
7. Average Base quality in normal samples is ≥ 25.
8. Distance of nearest InDel is > 25.
9. The proportion of soft-clipped reads is ≤ 0.25.

D.5 Parameters for Alternative Methods in Exome Sequence Data

VarScan2 (v2.3.9): –min-var-freq 0.01 –min-coverage 10 –min-coverage-normal 10 –min-coverage-tumor 10 –somatic-p-value 0.5.

Strelka (v1.0.14): isSkipDepthFilters = 1, and extraStrelkaArguments = –ignore-conflicting-read-names is set on the default setting.

Mutect (v1.1.4): –minimum_mutation_cell_fraction 0.01

OVarCall (v0.1.1): `https://github.com/takumorizo/OHVarfinDer/tree/master/utils/experiments/OVarCall/exome`

D.6 Parameters for Alternative Methods in Whole Genome Data

VarScan2 (v2.3.9): –min-var-freq 0.01 –min-coverage 10 –min-coverage-normal 10 –min-coverage-tumor 10 –somatic-p-value 0.5.

Strelka (v1.0.14): is set on the default setting.

Mutect (v1.1.4): –minimum_mutation_cell_fraction 0.01

OVarCall (v0.1.1): `https://github.com/takumorizo/OHVarfinDer/tree/master/utils/experiments/OVarCall/whole`

D.7 Comparison of running time of OHVarfinDer

The following table is a result of running time comparison of exome sequence data of RCC102, which is used in the performance evaluation of OVarCall.

Table 4: Performance comparison summary

Methods	Time	Memory
OVarCall	4:23:12	1.5G
OHVarfinDer	13:41:46	0.4G
HapMuC	15:35:37	0.4G
Strelka	3:23:38	0.7G
MuTect	7:0:36	3.0G
VarScan2	2:2:2	2.9G

D.8 Basic Information about original exome sequence data

		F3840			F2816			
OriginalReads	MappedReads	PE Reads	PEOVerlapReads	Overlap (%)	PE Reads	PEOVerlapReads	Overlap (%)	Sample
1.04E08	9.96E07	4.83E07	1.99E07	41.2	9.89E07	4.23E07	42.8	RCC102N
1.55E08	1.47E08	8.96E07	3.48E07	38.8	1.46E08	6.06E07	41.5	RCC102T
6.67E07	6.44E07	4.13E07	1.86E07	45.0	6.40E07	2.97E07	46.4	RCC104N
6.52E07	6.28E07	3.98E07	1.73E07	43.5	6.24E07	2.81E07	45.0	RCC104T
7.56E07	7.26E07	4.29E07	1.75E07	40.8	7.22E07	3.03E07	42.0	RCC161N
1.02E08	9.68E07	4.78E07	2.07E07	43.3	9.63E07	4.29E07	44.5	RCC161T
1.39E08	1.32E08	1.06E08	1.35E07	12.7	1.32E08	1.74E07	13.2	RCC163N
7.52E07	7.28E07	6.29E07	6.41E06	10.2	7.27E07	7.51E06	10.3	RCC163T
7.76E07	7.59E07	7.21E07	4.50E07	62.4	7.51E07	4.69E07	62.5	RCC172N
7.60E07	7.44E07	6.95E07	4.15E07	59.7	7.38E07	4.41E07	59.8	RCC172T
7.81E07	7.51E07	4.57E07	1.85E07	40.5	7.47E07	3.10E07	41.5	RCC179N
1.50E08	1.42E08	8.95E07	2.99E07	33.4	1.42E08	4.96E07	34.9	RCC179T
1.18E08	1.13E08	8.99E07	1.10E07	12.2	1.13E08	1.43E07	12.7	RCC183N
9.75E07	9.50E07	7.01E07	9.31E06	13.3	9.48E07	1.30E07	13.7	RCC183T
8.68E07	8.36E07	5.59E07	2.11E07	37.7	8.32E07	3.24E07	38.9	RCC185N
8.67E07	8.39E07	4.72E07	1.54E07	32.6	8.35E07	2.87E07	34.4	RCC185T
8.41E07	8.09E07	6.52E07	7.00E06	10.7	8.08E07	8.90E06	11.0	RCC197N
7.10E07	6.85E07	5.92E07	7.06E06	11.9	6.84E07	8.32E06	12.2	RCC197T
8.56E07	8.32E07	6.51E07	5.16E06	7.93	8.30E07	6.74E06	8.12	RCC201N
6.03E07	5.86E07	5.05E07	1.71E07	33.9	5.83E07	2.00E07	34.3	RCC201T
1.13E08	1.07E08	6.84E07	1.33E07	19.4	1.07E08	2.04E07	19.1	RCC252N
2.06E08	1.92E08	1.00E08	2.44E07	24.4	1.92E08	4.44E07	23.1	RCC252T
1.48E08	1.40E08	8.78E07	1.81E07	20.6	1.40E08	3.02E07	21.6	RCC295N
1.46E08	1.38E08	8.13E07	1.94E07	23.9	1.38E08	3.24E07	23.5	RCC295T
7.39E07	7.05E07	4.27E07	1.87E07	43.8	7.01E07	3.17E07	45.2	RCC297N
1.21E08	1.14E08	5.14E07	2.21E07	43.0	1.14E08	5.10E07	44.7	RCC297T
1.86E08	1.75E08	1.02E08	2.05E07	20.1	1.74E08	3.72E07	21.4	RCC312N
9.57E07	9.09E07	6.30E07	1.28E07	20.3	9.07E07	1.81E07	20.0	RCC312T
1.20E08	1.13E08	5.91E07	2.56E07	43.3	1.12E08	5.10E07	45.5	RCC31N
8.79E07	8.38E07	5.92E07	2.37E07	40.0	8.33E07	3.46E07	41.5	RCC31T
9.49E07	9.16E07	6.88E07	7.65E06	11.1	9.14E07	1.05E07	11.5	RCC324N
9.85E07	9.48E07	7.60E07	1.59E07	20.9	9.45E07	1.96E07	20.7	RCC324T
1.08E08	1.04E08	7.91E07	1.42E07	18.0	1.04E08	1.77E07	17.0	RCC34N
1.90E08	1.81E08	1.16E08	2.55E07	22.0	1.80E08	3.78E07	21.0	RCC34T
1.13E08	1.09E08	8.69E07	2.87E07	33.0	1.08E08	3.63E07	33.6	RCC58N
6.13E07	5.95E07	4.40E07	1.21E07	27.5	5.93E07	1.67E07	28.2	RCC58T
1.34E08	1.28E08	8.01E07	3.04E07	38.0	1.27E08	5.01E07	39.4	RCC88N
1.19E08	1.14E08	7.59E07	3.14E07	41.4	1.13E08	4.84E07	42.8	RCC88T
1.14E08	1.09E08	7.21E07	2.74E07	38.0	1.09E08	4.29E07	39.4	RCC95N
6.66E07	6.42E07	4.06E07	1.69E07	41.6	6.37E07	2.74E07	43.0	RCC95T

Figure 4: This table shows the basic informations of exome sequence data. We counted mapped reads by checking that the sam flag satisfies the following condition: read is mapped in proper pair. We counted paired-end reads by checking the sam flag satisfies the following condition: read is paired, read is mapped in proper pair, read reverse strand, after applying sam flag filter of -F 3840 or -F 2816. Sequence alignment was done by bwa 0.7.8.

D.9 Basic Information about original whole genome sequence data

		F3840			F2816			
OriginalReads	MappedReads	PE Reads	PEOVerlapReads	Overlap (%)	PE Reads	PEOVerlapReads	Overlap (%)	Sample
4.31E08	4.15E08	3.97E08	5.59E07	1.41E-01	4.11E08	5.65E07	1.37E-01	HCC1143_n100
4.17E08	3.61E08	3.45E08	1.94E06	5.63E-03	3.58E08	2.01E06	5.61E-03	HCC1143_n40t60
4.14E08	3.65E08	3.51E08	1.05E07	2.99E-02	3.62E08	1.06E07	2.93E-02	HCC1143_n60t40
4.49E08	3.82E08	3.61E08	4.13E05	1.15E-03	3.79E08	4.50E05	1.19E-03	HCC1954_n100
4.19E08	3.64E08	3.48E08	3.40E05	9.77E-04	3.61E08	3.81E05	1.06E-03	HCC1954_n40t60
4.21E08	3.63E08	3.48E08	3.65E05	1.05E-03	3.60E08	4.05E05	1.13E-03	HCC1954_n60t40

Figure 5: This table shows the basic informations of whole genome sequence data. We counted mapped reads by checking that the sam flag satisfies the following condition: read is mapped in proper pair. We counted paired-end reads by checking the sam flag satisfies the following condition: read is paired, read is mapped in proper pair, read reverse strand, after applying sam flag filter of -F 3840 or -F 2816. Sequence alignment was done by bwa 0.7.8.