JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 26, Number 9, 2019
Mary Ann Liebert, Inc.
Pp. 923–937
DOI: 10.1089/cmb.2018.0224

Other Research Articles

ALPHLARD-NT: Bayesian Method for Human Leukocyte Antigen Genotyping and Mutation Calling through Simultaneous Analysis of Normal and Tumor Whole-Genome Sequence Data

SHUTO HAYASHI[1], TAKUYA MORIYAMA[1], RUI YAMAGUCHI[1], SHINICHI MIZUNO[2], MITSUHIRO KOMURA[1], SATORU MIYANO[1], HIDEWAKI NAKAGAWA[3], and SEIYA IMOTO[4]

ABSTRACT

Human leukocyte antigen (HLA) genes provide useful information on the relationship between cancer and the immune system. Despite the ease of obtaining these data through next-generation sequencing methods, interpretation of these relationships remains challenging owing to the complexity of HLA genes. To resolve this issue, we developed a Bayesian method, ALPHLARD-NT, to identify HLA germline and somatic mutations as well as HLA genotypes from whole-exome sequencing (WES) and whole-genome sequencing (WGS) data. ALPHLARD-NT showed 99.2% accuracy for WGS-based HLA genotyping and detected five HLA somatic mutations in 25 colon cancer cases. In addition, ALPHLARD-NT identified 88 HLA somatic mutations, including recurrent mutations and a novel HLA-B type, from WES data of 343 colon adenocarcinoma cases. These results demonstrate the potential of ALPHLARD-NT for conducting an accurate analysis of HLA genes even from low-coverage data sets. This method can become an essential tool for comprehensive analyses of HLA genes from WES and WGS data, helping to advance understanding of immune regulation in cancer as well as providing guidance for novel immunotherapy strategies.

Keywords: Bayesian model, HLA genotyping, HLA mutation calling, whole-exome sequencing, whole-genome sequencing.

1. INTRODUCTION

HUMAN LEUKOCYTE ANTIGEN (HLA) GENES are essential components of the immune system, which present peptides to immune cells to facilitate recognition of nonself antigens. HLA genes must be

[1]Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan.
[2]Department of Health Sciences, Faculty of Medical Sciences, Kyushu University, Fukuoka, Japan.
[3]RIKEN Center for Integrative Medical Sciences, Tokyo, Japan.
[4]Health Intelligence Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan.

highly polymorphic to effectively carry out this function, with many types or alleles recognized, resulting in high individual variation in immune responses. Therefore, HLA genotyping, in which the specific pair of HLA types is identified for each HLA locus, is essential to understand the immune system. Recently, the interaction between cancer and the immune system has attracted attention (Grivennikov et al., 2010; Schreiber et al., 2011; Kreiter et al., 2015; Rooney et al., 2015; Marty et al., 2017), and somatic mutations in HLA genes have been shown to accumulate in specific cancer types (The Cancer Genome Atlas Research Network, 2014; Testoni et al., 2015; The Cancer Genome Atlas Network, 2015; Giannakis et al., 2016; McGranahan et al., 2017). Therefore, HLA genotyping can further help to understand the link between cancer and immunity, which would benefit personalized medicine.

There are several approaches currently available for HLA genotyping. Conventional approaches use polymerase chain reaction-based methods with sequence-specific oligonucleotides (Saiki et al., 1986), sequence-specific primers (Olerup and Zetterquist, 1992), and sequence-based typing (Santamaria et al., 1992); however, these methods are time consuming and labor intensive, and can only provide information on targeted HLA genes. New methods for HLA genotyping have been developed more recently with advances in molecular techniques, including whole-exome sequencing (WES), whole-genome sequencing (WGS), and RNA sequencing (Boegel et al., 2012; Warren et al., 2012; Kim and Pourmand 2013; Liu et al., 2013; Bai et al., 2014; Szolek et al., 2014; Nariai et al., 2015; Shukla et al., 2015; Dilthey et al., 2016; Xie et al., 2017; Hayashi et al., 2018; Lee and Kingsford, 2018). With these methods, information of both somatic mutations and HLA genotypes can be obtained from the entire sequence, which can facilitate investigations on the relationship between cancer and the immune system. In particular, methods that can specifically call germline or somatic mutations in HLA genes (Shukla et al., 2015; Hayashi et al., 2018; Lee and Kingsford, 2018) are valuable, since these mutations have potential to change immune responses, including tumor immune escape. However, the low coverage of WGS data makes it challenging to detect HLA germline and somatic mutations.

Previously, we developed a Bayesian model, called ALPHLARD (Hayashi et al., 2018), which identifies HLA genotypes and germline mutations from WGS data. ALPHLARD can also call HLA somatic mutations by comparing HLA sequences determined from normal and tumor samples. However, the specificity of the HLA somatic mutation calling is insufficient because ALPHLARD conducts the analyses of normal and tumor samples independently. To resolve this issue, we extended ALPHLARD to construct a new model named ALPHLARD-NT for accurately identifying both HLA germline and somatic mutations as well as HLA genotypes from WGS data. ALPHLARD-NT was validated from WES and WGS data sets from 343 and 25 colon cancer samples, respectively, which demonstrated its good performance in HLA genotyping, along with the ability to call HLA germline and somatic mutations, even from low-coverage data.

2. METHODS

2.1. Human leukocyte antigen reference data

We used the IPD-IMGT/HLA Database (Robinson et al., 2015) as HLA reference sequences in our method. Since the database provides incomplete sequences for most HLA types, we replaced the unknown bases with those of the most similar HLA type. To this end, similarity was determined by measuring the hamming distance in multiple sequence alignments (MSAs) across HLA types obtained from the IPD-IMGT/HLA Database. We used the Allele Frequency Net Database (González-Galarza et al., 2015) for prior information on HLA type frequencies.

2.2. Human leukocyte antigen read filtering and realignment

Filtering of HLA reads must be carefully performed for various reasons. First, it is insufficient to use only a human genome reference such as GRCh37 or GRCh38 owing to the high polymorphism of HLA genes. Therefore, a specific HLA database is required, such as the IPD-IMGT/HLA Database. Second, HLA genes and pseudogenes are paralogs and are, therefore, quite similar. Hence, when performing HLA genotyping, it is essential to distinguish reads from an HLA gene of interest from those of other HLA genes and pseudogenes.

In our HLA genotyping pipeline, a BAM file whose reference is the human genome is used as input data. First, sequence reads in the BAM file are filtered by extracting the HLA region, which is defined by chr6:28,477,797–33,448,354 for GRCh37 and chr6:28,510,120–33,480,577 for GRCh38, and covers the HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, and -DRB1 genes. Next, the extracted reads are mapped

to all HLA reference sequences using BWA-MEM (version 0.7.17) with the option to obtain information on all identified alignments. Each read is classified based on whether or not the HLA genes produced the read, and if so, which specific gene was involved. This classification is made using alignment scores, which we call HLA read scores (HR scores), and are calculated as follows. Let x_i be the i^{th} read pair that consists of two single reads $x_{i,0}$ and $x_{i,1}$. In the case of single-end sequence data, x_i consists of one read, $x_{i,0}$. In addition, t_k is defined as the k^{th} HLA type. If the read $x_{i,j}$ is unmapped to the HLA type t_k, then the HR score $s_{i,j,k}$ for $x_{i,j}$ and t_k is $-\infty$. Otherwise, $\tilde{x}_{i,j,k}$ and $\tilde{t}_{i,j,k}$ are the aligned sequences of $x_{i,j}$ and t_k, while $\tilde{x}_{i,j,k,n}$ and $\tilde{t}_{i,j,k,n}$ are the n^{th} bases or gaps of $\tilde{x}_{i,j,k}$ and $\tilde{t}_{i,j,k}$, respectively. Moreover, the mismatch probability $\tilde{q}_{i,j,k,n}$ of $\tilde{x}_{i,j,k,n}$ and $\tilde{t}_{i,j,k,n}$ can be calculated by

$$\tilde{q}_{i,j,k,n} = 10^{-\frac{\tilde{b}_{i,j,k,n}}{10}},$$

where $\tilde{b}_{i,j,k,n}$ is the Phred base quality of $\tilde{x}_{i,j,k,n}$. Using the aforementioned definitions, the HR score $s_{i,j,k}$ is given by

$$s_{i,j,k} = \sum_n (s^{(\mathtt{r})}_{i,j,k,n} + s^{(\mathtt{p})}_{i,j,k,n}),$$

where

$$s^{(\mathtt{r})}_{i,j,k,n} = \begin{cases} \alpha^{(\mathtt{r})} & (\text{if } \tilde{x}_{i,j,k,n} \in B^{(\mathtt{N})}) \\ 0 & (\text{if } \tilde{x}_{i,j,k,n} = -) \end{cases},$$

$$s^{(\mathtt{p})}_{i,j,k,n} = \begin{cases} 0 & (\text{if } \tilde{x}_{i,j,k,n},\ \tilde{t}_{i,j,k,n} \in B \text{ and } \tilde{x}_{i,j,k,n} = \tilde{t}_{i,j,k,n}) \\ \log\left(\frac{\tilde{q}_{i,j,k,n}}{3}\right) & (\text{if } \tilde{x}_{i,j,k,n},\ \tilde{t}_{i,j,k,n} \in B \text{ and } \tilde{x}_{i,j,k,n} \neq \tilde{t}_{i,j,k,n}) \\ \alpha^{(\mathtt{d},\mathtt{o})} & (\text{if } \tilde{x}_{i,j,k,n} = - \text{ and } \tilde{x}_{i,j,k,n-1} \neq -) \\ \alpha^{(\mathtt{d},\mathtt{e})} & (\text{if } \tilde{x}_{i,j,k,n} = - \text{ and } \tilde{x}_{i,j,k,n-1} = -) \\ \alpha^{(\mathtt{i},\mathtt{o})} & (\text{if } \tilde{t}_{i,j,k,n} = - \text{ and } \tilde{t}_{i,j,k,n-1} \neq -) \\ \alpha^{(\mathtt{i},\mathtt{e})} & (\text{if } \tilde{t}_{i,j,k,n} = - \text{ and } \tilde{t}_{i,j,k,n-1} = -) \\ \alpha^{(\mathtt{N})} & (\text{if } \tilde{x}_{i,j,k,n} = \mathtt{N} \text{ and } \tilde{t}_{i,j,k,n} \in B^{(\mathtt{N})} \text{ or } \tilde{x}_{i,j,k,n} \in B^{(\mathtt{N})} \text{ and } \tilde{t}_{i,j,k,n} = \mathtt{N}) \end{cases}$$

Here, $B = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ and $B^{(\mathtt{N})} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}, \mathtt{N}\}$. $s^{(\mathtt{r})}_{i,j,k,n}$ is a reward for the length of the read, and $\alpha^{(\mathtt{r})}$ is a positive hyperparameter for one base. By contrast, $s^{(\mathtt{p})}_{i,j,k,n}$ is a penalty for mismatches between the read and the HLA type, and $\alpha^{(\mathtt{d},\mathtt{o})}$, $\alpha^{(\mathtt{d},\mathtt{e})}$, $\alpha^{(\mathtt{i},\mathtt{o})}$, $\alpha^{(\mathtt{i},\mathtt{e})}$, and $\alpha^{(\mathtt{N})}$ are negative hyperparameters for deletion opening, deletion extension, insertion opening, insertion extension, and an unknown base N in the read or the HLA type, respectively.

Then, for each read pair x_i and each HLA locus l, the score $s^*_{i,l}$ is defined by

$$s^*_{i,l} = \sum_j \max_{k:t_k \in T_l} s_{i,j,k},$$

where T_l is a set of HLA types of the HLA locus l. When x_i is a paired-end read, it is used for genotyping the HLA locus l if the following two criteria are satisfied:

$$s^*_{i,l} > \theta^{(\mathtt{p},\mathtt{s})},$$

$$s^*_{i,l} - \max_{l' \neq l} s^*_{i,l'} > \theta^{(\mathtt{p},\mathtt{d})},$$

Here, $\theta^{(\mathtt{p},\mathtt{s})}$ is a hyperparameter of a threshold for the maximum HR score of the locus and $\theta^{(\mathtt{p},\mathtt{d})}$ is a hyperparameter of a threshold for the difference between the maximum HR scores of the locus and other loci. However, if x_i is a single-ended read, different thresholds are used; in other words, x_i is used for genotyping the HLA locus l if

$$s^*_{i,l} > \theta^{(\mathtt{s},\mathtt{s})},$$

$$s^*_{i,l} - \max_{l' \neq l} s^*_{i,l'} > \theta^{(\mathtt{s},\mathtt{d})}.$$

The former criterion is necessary to collect reads that are likely to be produced by the locus, whereas the latter criterion is needed to exclude reads that might be produced by other loci.

Next, all of the read pairs that satisfy the conditions are realigned to the MSAs of the HLA types of the HLA locus l. Realignment of the read $x_{i,j}$ is performed using the best HLA type whose index is given by

$$k^* = \arg\max_{k:t_k \in T_l} s_{i,j,k},$$

and the realigned read $\hat{x}_{i,j}$ is obtained by aligning $x_{i,j}$ to the MSA $\hat{t}_{k^*}$ of the HLA type t_{k^*} to match the alignment $(\tilde{x}_{i,j,k^*}, \tilde{t}_{i,j,k^*})$. This is done by simply translating the positions of bases and gaps in $\tilde{t}_{i,j,k^*}$ into those in $\hat{t}_{k^*}$.

2.3. Bayesian model for human leukocyte antigen analysis

We applied a Bayesian model for HLA genotyping and HLA somatic mutation detection, with basically the same structure as our previous method (Hayashi et al., 2018) except for some additional parameters. Figure 1 shows the graphical model. Hereafter, we suppose that the sequence reads are paired-ended for simplicity, and the model for single-ended sequence reads is the same except that the reads are unpaired.

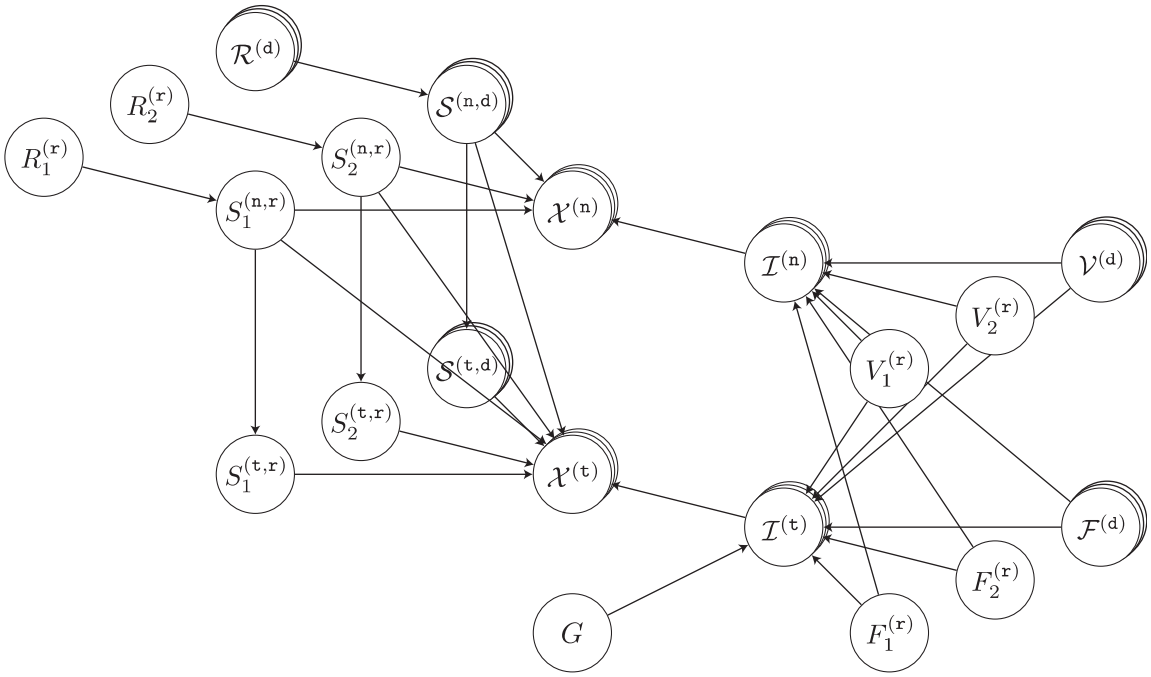
Input data of the model include both the normal and tumor realigned reads. Let $x_i^{(\mathtt{n})} = (x_{i,0}^{(\mathtt{n})}, x_{i,1}^{(\mathtt{n})})$ be the i^{th} normal realigned read pair, and $x_i^{(\mathtt{t})} = (x_{i,0}^{(\mathtt{t})}, x_{i,1}^{(\mathtt{t})})$ be the i^{th} tumor realigned read pair, where $\mathtt{n}$ and $\mathtt{t}$ indicate parameters for the normal and tumor sample, respectively. For each $s \in \{\mathtt{n}, \mathtt{t}\}$, we define $x_{i,j,n}^{(s)}$ as the n^{th} base of $x_{i,j}^{(s)}$, and $q_{i,j,n}^{(s)}$ as the mismatch probability of $x_{i,j,n}^{(s)}$. Note that the first position of each realigned read is not the beginning of the read but rather that of the MSAs, and $x_{i,j,n}^{(s)}$ and $q_{i,j,n}^{(s)}$ are undefined if the n^{th} position is not covered by the read. We define $r_{i,j}^{(s)}$ as a set of positions covered by the read $x_{i,j}^{(s)}$ and $r_i^{(s)}$ as $(r_{i,0}^{(s)}, r_{i,1}^{(s)})$.

We denote HLA types of the sample by $R_1^{(\mathtt{r})}$ and $R_2^{(\mathtt{r})}$, normal HLA sequences by $S_1^{(\mathtt{n},\mathtt{r})}$ and $S_2^{(\mathtt{n},\mathtt{r})}$, and tumor HLA sequences by $S_1^{(\mathtt{t},\mathtt{r})}$ and $S_2^{(\mathtt{t},\mathtt{r})}$. Here, the sequences of $R_1^{(\mathtt{r})}$ and $R_2^{(\mathtt{r})}$ are the MSAs of the HLA types. $S_1^{(\mathtt{n},\mathtt{r})}$ and $S_2^{(\mathtt{n},\mathtt{r})}$ are used to consider germline variants in $R_1^{(\mathtt{r})}$ and $R_2^{(\mathtt{r})}$, and $S_1^{(\mathtt{t},\mathtt{r})}$ and $S_2^{(\mathtt{t},\mathtt{r})}$ are used to reflect somatic mutations. We also introduce decoy HLA types $R_1^{(\mathtt{d})}, \ldots, R_{\nu^{(\mathtt{d})}}^{(\mathtt{d})}$, decoy normal HLA sequences $S_1^{(n,d)}, \ldots, S_{\nu^{(\mathtt{d})}}^{(\mathtt{n},\mathtt{d})}$, and decoy tumor HLA sequences $S_1^{(\mathtt{t},\mathtt{d})}, \ldots, S_{\nu^{(\mathtt{d})}}^{(\mathtt{t},\mathtt{d})}$, where $\nu^{(\mathtt{d})}$ is a hyperparameter of the number of the decoy parameters. These parameters are essential to make a robust inference, because their presence can reduce the influence of misclassified reads at the previous filtering step that were actually produced by other HLA genes or pseudogenes. For convenience, we sometimes use $(R_1, R_2, R_3, \ldots, R_{\nu^{(\mathtt{d})}+2})$, $(S_1^{(\mathtt{n})}, S_2^{(\mathtt{n})}, S_3^{(\mathtt{n})}, \ldots, S_{\nu^{(\mathtt{d})}+2}^{(\mathtt{n})})$, and $(S_1^{(\mathtt{t})}, S_2^{(\mathtt{t})}, S_3^{(\mathtt{t})}, \ldots, S_{\nu^{(\mathtt{d})}+2}^{(\mathtt{t})})$ instead of $(R_1^{(\mathtt{r})}, R_2^{(\mathtt{r})}, R_1^{(\mathtt{d})}, \ldots, R_{\nu^{(\mathtt{d})}}^{(\mathtt{d})})$, $(S_1^{(\mathtt{n},\mathtt{r})}, S_2^{(\mathtt{n},\mathtt{r})}, S_1^{(\mathtt{n},\mathtt{d})}, \ldots, S_{\nu^{(\mathtt{d})}}^{(\mathtt{n},\mathtt{d})})$, and $(S_1^{(\mathtt{t},\mathtt{r})}, S_2^{(\mathtt{t},\mathtt{r})}, S_1^{(\mathtt{t},\mathtt{d})}, \ldots, S_{\nu^{(\mathtt{d})}}^{(\mathtt{t},\mathtt{d})})$, respectively. In addition, in some cases, $(S_1, \ldots, S_{2\nu^{(\mathtt{d})}+4})$ is used instead of $(S_1^{(\mathtt{n})}, \ldots, S_{\nu^{(\mathtt{d})}+2}^{(\mathtt{n})}, S_1^{(\mathtt{t})}, \ldots, S_{\nu^{(\mathtt{d})}+2}^{(\mathtt{t})})$. Similar to the notation for read pairs, $R_{m,n}$ and $S_{m,n}$ are defined as the n^{th} base of R_m and S_m, respectively.

Next, let $I_i^{(\mathtt{n})}$ and $I_i^{(\mathtt{t})}$ be parameters that indicate the specific HLA sequence that produced $x_i^{(\mathtt{n})}$ and $x_i^{(\mathtt{t})}$, respectively. In other words, $I_i^{(s)} = m$ means that $x_i^{(s)}$ was produced by S_m. Note that $I_i^{(\mathtt{n})} \in \{1, \ldots, \nu^{(\mathtt{d})}+2\}$ because tumor HLA sequences cannot produce normal sequence reads, and that $I_i^{(\mathtt{t})} \in \{1, \ldots, 2\nu^{(\mathtt{d})}+4\}$ because the tumor sample might also contain normal cells. $I_i^{(s)}$ is independently generated from a distribution governed by $F_1^{(\mathtt{r})}, F_2^{(\mathtt{r})}, F_1^{(\mathtt{d})}, \ldots, F_{\nu^{(\mathtt{d})}}^{(\mathtt{d})}$, G, and $V_1^{(\mathtt{r})}, V_2^{(\mathtt{r})}, V_1^{(\mathtt{d})}, \ldots, V_{\nu^{(\mathtt{d})}}^{(\mathtt{d})}$. Again, we sometimes use convenient notations of $(F_1, F_2, F_3, \ldots, F_{\nu^{(\mathtt{d})}+2})$ and $(V_1, V_2, V_3, \ldots, V_{\nu^{(\mathtt{d})}+2})$ instead of $(F_1^{(\mathtt{r})}, F_2^{(\mathtt{r})}, F_1^{(\mathtt{d})}, \ldots, F_{\nu^{(\mathtt{d})}}^{(\mathtt{d})})$, and $(V_1^{(\mathtt{r})}, V_2^{(\mathtt{r})}, V_1^{(\mathtt{d})}, \ldots, V_{\nu^{(\mathtt{d})}}^{(\mathtt{d})})$. F_m is a positive real parameter that expresses the likelihood that a read is produced by $S_m^{(\mathtt{n})}$ and $S_m^{(\mathtt{t})}$. G is also a positive real parameter and expresses the ratio of normal cells contained in the tumor sample. V_m is a tuple $(V_{m,1}, \ldots, V_{m,N})$, where N is the length of MSAs and $V_{m,n}$ is a parameter of 0 or 1, which indicates whether $S_{m,n}^{(\mathtt{n})}$ and $S_{m,n}^{(\mathtt{t})}$ are valid, as described in more detail hereunder.

The posterior probability of the parameters is given by

$$\begin{aligned} p(\mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})} | \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})}) \\ \propto p(\mathcal{X}^{(\mathtt{n})} | \mathcal{S}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{n})}) p(\mathcal{X}^{(\mathtt{t})} | \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{t})}) \\ \times p(\mathcal{S}^{(\mathtt{t})} | \mathcal{S}^{(\mathtt{n})}) p(\mathcal{S}^{(\mathtt{n})} | \mathcal{R}) p(\mathcal{R}) \\ \times p(\mathcal{I}^{(\mathtt{n})} | \mathcal{F}, \mathcal{V}) p(\mathcal{I}^{(\mathtt{t})} | \mathcal{F}, G, \mathcal{V}) p(\mathcal{F}) p(G) p(\mathcal{V}), \end{aligned}$$



$R_1^{(\mathrm{r})}$, $R_2^{(\mathrm{r})}$: HLA types
$\mathcal{R}^{(\mathrm{d})}$: decoy HLA types
$S_1^{(\mathrm{n,r})}$, $S_2^{(\mathrm{n,r})}$: normal HLA sequences
$\mathcal{S}^{(\mathrm{n,d})}$: decoy normal HLA sequences
$S_1^{(\mathrm{t,r})}$, $S_2^{(\mathrm{t,r})}$: tumor HLA sequences
$\mathcal{S}^{(\mathrm{t,d})}$: decoy tumor HLA sequences
$\mathcal{X}^{(\mathrm{n})}$: normal realigned read pairs
$\mathcal{X}^{(\mathrm{t})}$: tumor realigned read pairs
$\mathcal{I}^{(\mathrm{n})}$: variables that indicate which HLA sequence produced each normal read pair
$\mathcal{I}^{(\mathrm{t})}$: variables that indicate which HLA sequence produced each tumor read pair
V_1, V_2, $\mathcal{V}^{(\mathrm{d})}$: variables that indicate whether each HLA sequence is valid
F_1, F_2, $\mathcal{F}^{(\mathrm{d})}$: variables that express how likely each HLA sequence is to produce read pairs
G: a variable that expresses the ratio of normal cells contained in the tumor sample

FIG. 1. Graphical representation of our method.

where $\mathcal{R} = (R_1, \ldots, R_{\nu^{(\mathrm{d})}+2})$, $\mathcal{S}^{(\mathrm{n})} = (S_1^{(\mathrm{n})}, \ldots, S_{\nu^{(\mathrm{d})}+2}^{(\mathrm{n})})$, $\mathcal{S}^{(\mathrm{t})} = (S_1^{(\mathrm{t})}, \ldots, S_{\nu^{(\mathrm{d})}+2}^{(\mathrm{t})})$, $\mathcal{F} = (F_1, \ldots, F_{\nu^{(\mathrm{d})}+2})$, $\mathcal{V} = (V_1, \ldots, V_{\nu^{(\mathrm{d})}+2})$, $\mathcal{I}^{(\mathrm{n})} = (I_1^{(\mathrm{n})}, I_2^{(\mathrm{n})}, \ldots)$, $\mathcal{I}^{(\mathrm{t})} = (I_1^{(\mathrm{t})}, I_2^{(\mathrm{t})}, \ldots)$, $\mathcal{X}^{(\mathrm{n})} = (x_1^{(\mathrm{n})}, x_2^{(\mathrm{n})}, \ldots)$, and $\mathcal{X}^{(\mathrm{t})} = (x_1^{(\mathrm{t})}, x_2^{(\mathrm{t})}, \ldots)$.

The likelihoods of sequence read pairs are given by

$$p(\mathcal{X}^{(\mathrm{n})} | \mathcal{S}^{(\mathrm{n})}, \mathcal{I}^{(\mathrm{n})}) = \prod_i \prod_j \prod_n p(x_{i,j,n}^{(\mathrm{n})} | S_{I_i^{(\mathrm{n})},n}),$$

$$p(\mathcal{X}^{(\mathrm{t})} | \mathcal{S}^{(\mathrm{n})}, \mathcal{S}^{(\mathrm{t})}, \mathcal{I}^{(\mathrm{t})}) = \prod_i \prod_j \prod_n p(x_{i,j,n}^{(\mathrm{t})} | S_{I_i^{(\mathrm{t})},n}),$$

where

$$p(x_{i,j,n}|S_{m,n} \in B) = \begin{cases} (1-\pi^{(\mathrm{e},\mathrm{d})})(1-\pi^{(\mathrm{e},\mathrm{N})})(1-p_{i,j,n}) & (\text{if } x_{i,j,n} = S_{m,n}) \\ (1-\pi^{(\mathrm{e},\mathrm{d})})(1-\pi^{(\mathrm{e},\mathrm{N})})\frac{p_{i,j,n}}{3} & (\text{if } x_{i,j,n} \in B \text{ and } x_{i,j,n} \neq S_{m,n}) \\ (1-\pi^{(\mathrm{e},\mathrm{d})})\pi^{(\mathrm{e},\mathrm{N})} & (\text{if } x_{i,j,n} = \mathrm{N}) \\ \pi^{(\mathrm{e},\mathrm{d})} & (\text{if } x_{i,j,n} = -) \end{cases},$$

$$p(x_{i,j,n}|S_{m,n} = -) = \begin{cases} \pi^{(\mathrm{e},\mathrm{i})}(1-\pi^{(\mathrm{e},\mathrm{N})})\frac{1}{4} & (\text{if } x_{i,j,n} \in B) \\ \pi^{(\mathrm{e},\mathrm{i})}\pi^{(\mathrm{e},\mathrm{N})} & (\text{if } x_{i,j,n} = \mathrm{N}) \\ 1-\pi^{(\mathrm{e},\mathrm{i})} & (\text{if } x_{i,j,n} = -) \end{cases},$$

$$p(x_{i,j,n}|S_{m,n} = \mathrm{N}) = \begin{cases} (1-\pi^{(\mathrm{e},\mathrm{N})})\frac{1}{5} & (\text{if } x_{i,j,n} \in B \text{ or } x_{i,j,n} -) \\ \pi^{(\mathrm{e},\mathrm{N})} & (\text{if } x_{i,j,n} = \mathrm{N}) \end{cases}$$

Here, $\pi^{(\mathrm{e},\mathrm{d})}$, $\pi^{(\mathrm{e},\mathrm{i})}$, and $\pi^{(\mathrm{e},\mathrm{N})}$ are hyperparameters of probabilities of a deletion error, insertion error, and N in a sequence read, respectively.

The prior probability of tumor HLA sequences is given by

$$p(\mathcal{S}^{(\mathrm{t})}|\mathcal{S}^{(\mathrm{n})}) = \prod_m \prod_n p(S^{(\mathrm{t})}_{m,n}|S^{(\mathrm{n})}_{m,n}),$$

where

$$p(S^{(\mathrm{t})}_{m,n}|S^{(\mathrm{n})}_{m,n} \in B) = \begin{cases} (1-\pi^{(\mathrm{s},\mathrm{N})})(1-\pi^{(\mathrm{s},\mathrm{d})})(1-\pi^{(\mathrm{s},\mathrm{s})}) & (\text{if } S^{(\mathrm{t})}_{m,n} = S^{(\mathrm{n})}_{m,n}) \\ (1-\pi^{(\mathrm{s},\mathrm{N})})(1-\pi^{(\mathrm{s},\mathrm{d})})\frac{\pi^{(\mathrm{s},\mathrm{s})}}{3} & (\text{if } S^{(\mathrm{t})}_{m,n} \in B \text{ and } S^{(\mathrm{t})}_{m,n} \neq S^{(\mathrm{n})}_{m,n}) \\ (1-\pi^{(\mathrm{s},\mathrm{N})})\pi^{(\mathrm{s},\mathrm{d})} & (\text{if } S^{(\mathrm{t})}_{m,n} = -) \\ \pi^{(\mathrm{s},\mathrm{N})} & (\text{if } S^{(\mathrm{t})}_{m,n} = \mathrm{N}) \end{cases},$$

$$p(S^{(\mathrm{t})}_{m,n}|S^{(\mathrm{n})}_{m,n} = -) = \begin{cases} (1-\pi^{(\mathrm{s},\mathrm{N})})\pi^{(\mathrm{s},\mathrm{i})}\frac{1}{4} & (\text{if } S^{(\mathrm{t})}_{m,n} \in B) \\ (1-\pi^{(\mathrm{s},\mathrm{N})})(1-\pi^{(\mathrm{s},\mathrm{i})}) & (\text{if } S^{(\mathrm{t})}_{m,n} = -) \\ \pi^{(\mathrm{s},\mathrm{N})} & (\text{if } S^{(\mathrm{t})}_{m,n} = \mathrm{N}) \end{cases},$$

$$p(S^{(\mathrm{t})}_{m,n}|S^{(\mathrm{n})}_{m,n} = \mathrm{N}) = \begin{cases} (1-\pi^{(\mathrm{s},\mathrm{N})})\frac{1}{5} & (\text{if } S^{(\mathrm{t})}_{m,n} \in B \text{ or } S^{(\mathrm{t})}_{m,n} = -) \\ \pi^{(\mathrm{s},\mathrm{N})} & (\text{if } S^{(\mathrm{t})}_{m,n} = \mathrm{N}) \end{cases}$$

Here, $\pi^{(\mathrm{s},\mathrm{s})}$, $\pi^{(\mathrm{s},\mathrm{d})}$, $\pi^{(\mathrm{s},\mathrm{i})}$, and $\pi^{(\mathrm{s},\mathrm{N})}$ are hyperparameters of probabilities of a somatic substitution, somatic deletion, somatic insertion, and N in a tumor HLA sequence, respectively.

The prior probability of normal HLA sequences is given by

$$p(\mathcal{S}^{(\mathrm{n})}|\mathcal{R}) = \left(\prod_m \prod_n p\left(S^{(\mathrm{n},\mathrm{r})}_{m,n}|R^{(\mathrm{r})}_{m,n}\right)\right)\left(\prod_m \prod_n p\left(S^{(\mathrm{n},\mathrm{d})}_{m,n}|R^{(\mathrm{d})}_{m,n}\right)\right),$$

where

$$p(S^{(\mathtt{n},\mathtt{r})}_{m,n}|R^{(\mathtt{r})}_{m,n} \in B, R^{(\mathtt{r})}_{m,n} \text{ is original})$$
$$= \begin{cases} (1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{d})})(1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{s})}) & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} = R^{(\mathtt{r})}_{m,n}) \\ (1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{d})})\frac{\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{s})}}{3} & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} \in B \text{ and } S^{(\mathtt{n},\mathtt{r})}_{m,n} \neq R^{(\mathtt{r})}_{m,n}) \\ (1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{N})})\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{d})} & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} = -) \\ \pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{N})} & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} = \mathtt{N}) \end{cases},$$

$$p(S^{(\mathtt{n},\mathtt{r})}_{m,n}|R^{(\mathtt{r})}_{m,n} = -, R^{(\mathtt{r})}_{m,n} \text{ is original})$$
$$= \begin{cases} (1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{N})})\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{i})}\frac{1}{4} & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} \in B) \\ (1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{i})}) & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} = -) \\ \pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{N})} & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} = \mathtt{N}) \end{cases},$$

$$p(S^{(\mathtt{n},\mathtt{r})}_{m,n}|R^{(\mathtt{r})}_{m,n} = \mathtt{N}, R^{(\mathtt{r})}_{m,n} \text{ is original})$$
$$= \begin{cases} (1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{N})})\frac{1}{5} & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} \in B \text{ or } S^{(\mathtt{n},\mathtt{r})}_{m,n} = -) \\ \pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{N})} & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} = \mathtt{N}) \end{cases},$$

$$p(S^{(\mathtt{n},\mathtt{r})}_{m,n}|R^{(\mathtt{r})}_{m,n} \in B, R^{(\mathtt{r})}_{m,n} \text{ is imputed})$$
$$= \begin{cases} (1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{d})})(1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{s})}) & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} = R^{(\mathtt{r})}_{m,n}) \\ (1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{d})})\frac{\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{s})}}{3} & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} \in B \text{ and } S^{(\mathtt{n},\mathtt{r})}_{m,n} \neq R^{(\mathtt{r})}_{m,n}) \\ (1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{N})})\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{d})} & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} = -) \\ \pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{N})} & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} = \mathtt{N}) \end{cases},$$

$$p(S^{(\mathtt{n},\mathtt{r})}_{m,n}|R^{(\mathtt{r})}_{m,n} = -, R^{(\mathtt{r})}_{m,n} \text{ is imputed})$$
$$= \begin{cases} (1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{N})})\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{i})}\frac{1}{4} & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} \in B) \\ (1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{i})}) & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} = -) \\ \pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{N})} & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} = \mathtt{N}) \end{cases},$$

$$p(S^{(\mathtt{n},\mathtt{r})}_{m,n}|R^{(\mathtt{r})}_{m,n} = \mathtt{N}, R^{(\mathtt{r})}_{m,n} \text{ is imputed})$$
$$= \begin{cases} (1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{N})})\frac{1}{5} & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} \in B \text{ or } S^{(\mathtt{n},\mathtt{r})}_{m,n} = -) \\ \pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{N})} & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} = \mathtt{N}) \end{cases},$$

$$p(S^{(\mathtt{n},\mathtt{d})}_{m,n}|R^{(\mathtt{d})}_{m,n} \in B, R^{(\mathtt{d})}_{m,n} \text{ is original})$$
$$= \begin{cases} (1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{d})})(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{s})}) & (\text{if } S^{(\mathtt{n},\mathtt{d})}_{m,n} = R^{(\mathtt{d})}_{m,n}) \\ (1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{d})})\frac{\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{s})}}{3} & (\text{if } S^{(\mathtt{n},\mathtt{d})}_{m,n} \in B \text{ and } S^{(\mathtt{n},\mathtt{d})}_{m,n} \neq R^{(\mathtt{d})}_{m,n}) \\ (1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{N})})\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{d})} & (\text{if } S^{(\mathtt{n},\mathtt{d})}_{m,n} = -) \\ \pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{N})} & (\text{if } S^{(\mathtt{n},\mathtt{d})}_{m,n} = \mathtt{N}) \end{cases},$$

$$p(S^{(\mathtt{n},\mathtt{d})}_{m,n}|R^{(\mathtt{d})}_{m,n} = -, R^{(\mathtt{d})}_{m,n} \text{ is original})$$
$$= \begin{cases} (1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{N})})\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{i})}\frac{1}{4} & (\text{if } S^{(\mathtt{n},\mathtt{d})}_{m,n} \in B) \\ (1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{i})}) & (\text{if } S^{(\mathtt{n},\mathtt{d})}_{m,n} = -) \\ \pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{N})} & (\text{if } S^{(\mathtt{n},\mathtt{d})}_{m,n} = \mathtt{N}) \end{cases},$$

$$p(S^{(\mathtt{n},\mathtt{d})}_{m,n}|R^{(\mathtt{d})}_{m,n} = \mathtt{N}, R^{(\mathtt{d})}_{m,n} \text{ is original})$$
$$= \begin{cases} (1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{N})})\frac{1}{5} & (\text{if } S^{(\mathtt{n},\mathtt{d})}_{m,n} \in B \text{ or } S^{(\mathtt{n},\mathtt{d})}_{m,n} = -) \\ \pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{N})} & (\text{if } S^{(\mathtt{n},\mathtt{d})}_{m,n} = \mathtt{N}) \end{cases},$$

$$p(S_{m,n}^{(\mathtt{n},\mathtt{d})}|R_{m,n}^{(\mathtt{d})} \in B, R_{m,n}^{(\mathtt{d})} \text{ is imputed})$$
$$= \begin{cases} (1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{d})})(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{s})}) & (\text{if } S_{m,n}^{(\mathtt{n},\mathtt{d})} = R_{m,n}^{(\mathtt{d})}) \\ (1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{d})})\frac{\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{s})}}{3} & (\text{if } S_{m,n}^{(\mathtt{n},\mathtt{d})} \in B \text{ and } S_{m,n}^{(\mathtt{n},\mathtt{d})} \neq R_{m,n}^{(\mathtt{d})}) \\ (1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})})\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{d})} & (\text{if } S_{m,n}^{(\mathtt{n},\mathtt{d})} = -) \\ \pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})} & (\text{if } S_{m,n}^{(\mathtt{n},\mathtt{d})} = \mathtt{N}) \end{cases},$$
$$p(S_{m,n}^{(\mathtt{n},\mathtt{d})}|R_{m,n}^{(\mathtt{d})} = -, R_{m,n}^{(\mathtt{d})} \text{ is imputed})$$
$$= \begin{cases} (1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})})\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{i})}\frac{1}{4} & (\text{if } S_{m,n}^{(\mathtt{n},\mathtt{d})} \in B) \\ (1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{i})}) & (\text{if } S_{m,n}^{(\mathtt{n},\mathtt{d})} = -) \\ \pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})} & (\text{if } S_{m,n}^{(\mathtt{n},\mathtt{d})} = \mathtt{N}) \end{cases},$$
$$p(S_{m,n}^{(\mathtt{n},\mathtt{d})}|R_{m,n}^{(\mathtt{d})} = \mathtt{N}, R_{m,n}^{(\mathtt{d})} \text{ is imputed})$$
$$= \begin{cases} (1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})})\frac{1}{5} & (\text{if } S_{m,n}^{(\mathtt{n},\mathtt{d})} \in B \text{ or } S_{m,n}^{(\mathtt{n},\mathtt{d})} = -) \\ \pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})} & (\text{if } S_{m,n}^{(\mathtt{n},\mathtt{d})} = \mathtt{N}) \end{cases}$$

Here, $\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{s})}$, $\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{d})}$, $\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{i})}$, and $\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{N})}$ are hyperparameters of probabilities of a germline substitution, germline deletion, germline insertion, and N, respectively, in a nondecoy normal HLA sequence at the position where the reference is an original base. The other hyperparameters are also defined in a similar way. The probabilities for an imputed reference base should be larger than those for an original base to reduce the influence of misimputation. In addition, the probabilities for a decoy normal HLA sequence should also be larger than those for a nondecoy normal HLA sequence to achieve robustness against misclassified reads.

The prior probability of HLA types is given by

$$p(\mathcal{R}) = \left(\prod_m p\left(R_m^{(\mathtt{r})}\right)\right)\left(\prod_m p\left(R_m^{(\mathtt{d})}\right)\right),$$

where

$$p(R_m^{(\mathtt{r})} = t) = p_t,$$
$$p(R_m^{(\mathtt{d})}) \propto 1.$$

Here, p_t is a prior probability of the HLA type t, which was calculated using the Allele Frequency Net Database.

The prior probability of normal indicator variables is given by

$$p(\mathcal{I}^{(\mathtt{n})}|\mathcal{F}, \mathcal{V}) = \prod_i p(I_i^{(\mathtt{n})}|\mathcal{F}, \mathcal{V}),$$

where

$$p(I_i^{(\mathtt{n})} = m|\mathcal{F}, \mathcal{V}) \propto \left(\max_{n \in \cup_j r_{i,j}^{(\mathtt{n})}} V_m\right) F_m.$$

This formula means that the read cannot be produced by an HLA sequence without a valid position covered by the read, which is controlled by $\mathcal{V}$. Similarly, the prior probability of tumor indicator variables is given by

$$p(\mathcal{I}^{(\mathtt{t})}|\mathcal{F}, G, \mathcal{V}) = \prod_i p(I_i^{(\mathtt{t})}|\mathcal{F}, G, \mathcal{V}),$$

where

$$p(I_i^{(\mathtt{t})} = m \in M^{(\mathtt{n})}|\mathcal{F}, G, \mathcal{V}) \propto \left(\max_{n \in \cup_j r_{i,j}^{(\mathtt{t})}} V_m\right) F_m G,$$

$$p(I_i^{(\mathtt{t})} = m \in M^{(\mathtt{t})} | \mathcal{F}, G, \mathcal{V}) \propto (\max_{n \in \cup_j r_{i,j}^{(\mathtt{t})}} V_{m-(\nu^{(\mathtt{d})}+2)}) F_{m-(\nu^{(\mathtt{d})}+2)},$$

$$M^{(\mathtt{n})} = \{1, \ldots, \nu^{(\mathtt{d})} + 2\},$$

$$M^{(\mathtt{t})} = \{\nu^{(\mathtt{d})} + 3, \ldots, 2\nu^{(\mathtt{d})} + 4\}$$

Note that $I_i^{(\mathtt{t})} \in M^{(\mathtt{n})}$ indicates that the read was derived from a normal cell, and $I_i^{(\mathtt{t})} \in M^{(\mathtt{t})}$ indicates that the read was derived from a tumor cell. Furthermore, matched normal-tumor HLA sequences $S_m^{(\mathtt{n})}$ and $S_m^{(\mathtt{t})}$ share V_m and F_m.

The prior probability of $\mathcal{F}$ is given by

$$p(\mathcal{F}) = \left(\prod_m p\left(F_m^{(\mathtt{r})}\right) \right) \left(\prod_m p\left(F_m^{(\mathtt{d})}\right) \right),$$

where

$$p(F_m^{(\mathtt{r})}) = \mathcal{LN}(F_m^{(\mathtt{r})} | \mu^{(\mathtt{f},\mathtt{r})}, (\sigma^{(\mathtt{f},\mathtt{r})})^2),$$
$$p(F_m^{(\mathtt{d})}) = \mathcal{LN}(F_m^{(\mathtt{d})} | \mu^{(\mathtt{f},\mathtt{d})}, (\sigma^{(\mathtt{f},\mathtt{d})})^2)$$

Here, $\mathcal{LN}$ is a log-normal distribution, $\mu^{(\mathtt{f},\mathtt{r})}$ and $(\sigma^{(\mathtt{f},\mathtt{r})})^2$ are hyperparameters of the mean and variance for the nondecoy parameters, and $\mu^{(\mathtt{f},\mathtt{d})}$ and $(\sigma^{(\mathtt{f},\mathtt{d})})^2$ are hyperparameters of the mean and variance for the decoy parameters. $\mu^{(\mathtt{f},\mathtt{d})}$ should be smaller than $\mu^{(\mathtt{f},\mathtt{r})}$ because sequence reads mapped to decoy HLA sequences should be removed at the filtering step.

The prior probability of G is given by

$$p(G) = \mathcal{LN}(G | \mu^{(\mathtt{g})}, (\sigma^{(\mathtt{g})})^2),$$

where $\mu^{(\mathtt{g})}$ and $(\sigma^{(\mathtt{g})})^2$ are hyperparameters of the mean and variance for normal contamination.

The prior probability of $\mathcal{V}$ is given by

$$p(\mathcal{V}) = \left(\prod_m \prod_n p\left(V_{m,n}^{(\mathtt{r})}\right) \right) \left(\prod_m \prod_n p\left(V_{m,n}^{(\mathtt{d})} | V_{m,n-1}^{(\mathtt{d})}\right) \right),$$

where

$$p(V_{m,n}^{(\mathtt{r})}) = \begin{cases} 0 & (\text{if } V_{m,n}^{(\mathtt{r})} = 0) \\ 1 & (\text{if } V_{m,n}^{(\mathtt{r})} = 1) \end{cases},$$

$$p(V_{m,n}^{(\mathtt{d})} | V_{m,n-1}^{(\mathtt{d})} = 0) = \begin{cases} 1 - \pi^{(\mathtt{v},\mathtt{o})} & (\text{if } V_{m,n}^{(\mathtt{d})} = 0) \\ \pi^{(\mathtt{v},\mathtt{o})} & (\text{if } V_{m,n}^{(\mathtt{d})} = 1) \end{cases},$$

$$p(V_{m,n}^{(\mathtt{d})} | V_{m,n-1}^{(\mathtt{d})} = 1) = \begin{cases} 1 - \pi^{(\mathtt{v},\mathtt{e})} & (\text{if } V_{m,n}^{(\mathtt{d})} = 0) \\ \pi^{(\mathtt{v},\mathtt{e})} & (\text{if } V_{m,n}^{(\mathtt{d})} = 1) \end{cases}$$

Here, $\pi^{(\mathtt{v},\mathtt{o})}$ and $\pi^{(\mathtt{v},\mathtt{e})}$ are hyperparameters of probabilities of a validity flag opening and a validity flag extension, respectively. Note that $V_{m,n}^{(\mathtt{r})}$ must always be 1.

2.4. Markov chain Monte Carlo-based parameter sampling

The parameters are sampled from the Bayesian model using Markov chain Monte Carlo. Gibbs sampling is primarily used to sample all parameters except for F_m and V_m.

A candidate parameter, F_m^*, is first sampled using the Metropolis–Hastings algorithm whose proposal distribution is given by

$$F_m^* \sim \mathcal{LN}(\log F_m, (\sigma_m^{(\mathtt{f},\mathtt{p})})^2),$$

where $(\sigma_m^{(\mathtt{f},\mathtt{p})})^2$ is a hyperparameter of the variance of the proposal distribution. The acceptance ratio r^* is calculated by

$$r^* = \frac{p(\mathcal{I}^{(\mathrm{n})}|\mathcal{F}^*, \mathcal{V})p(\mathcal{I}^{(\mathrm{t})}|\mathcal{F}^*, \mathcal{V})p(F_m^*)}{p(\mathcal{I}^{(\mathrm{n})}|\mathcal{F}, \mathcal{V})p(\mathcal{I}^{(\mathrm{t})}|\mathcal{F}, \mathcal{V})p(F_m)},$$

where $\mathcal{F}^* = (F_1, \ldots, F_{m-1}, F_m^*, F_{m+1}, \ldots, F_{\nu^{(\mathrm{d})}+2})$. A candidate parameter, V_m^*, is sampled using the Metropolis–Hastings algorithm whose proposal distribution is analogous to the Wolff algorithm (Wolff, 1989), which is used for sampling of the Ising model. V_m^* is generated by Algorithm 1. Then, $\mathcal{I}^{(\mathrm{n})*}$ and $\mathcal{I}^{(\mathrm{t})*}$ are also sampled using Gibbs sampling given V_m^*. The acceptance ratio r^* is calculated by

$$\begin{aligned} r^* = & \frac{\prod_{n=1}^{r+1} p(V_{m,n}^*|V_{m,n-1}^*)}{\prod_{n=1}^{r+1} p(V_{m,n}|V_{m,n-1})} \\ & \times \frac{(\pi_v^{(\mathrm{v},\mathrm{p})})^{r-l}(1-\pi_v^{(\mathrm{v},\mathrm{p})})^{[l\neq 1\wedge V_{l-1}\neq v]+[r\neq N\wedge V_{r+1}\neq v]}}{(\pi_{1-v}^{(\mathrm{v},\mathrm{p})})^{r-l}(1-\pi_{1-v}^{(\mathrm{v},\mathrm{p})})^{[l\neq 1\wedge V_{l-1}\neq v]+[r\neq N\wedge V_{r+1}\neq v]}} \\ & \times \frac{p(\mathcal{X}^{(\mathrm{n})}|\mathcal{S}^{(\mathrm{n})}, \mathcal{V}^*)p(\mathcal{X}^{(\mathrm{t})}|\mathcal{S}^{(\mathrm{n})}, \mathcal{S}^{(\mathrm{t})}, \mathcal{V}^*)}{p(\mathcal{X}^{(\mathrm{n})}|\mathcal{S}^{(\mathrm{n})}, \mathcal{V})p(\mathcal{X}^{(\mathrm{t})}|\mathcal{S}^{(\mathrm{n})}, \mathcal{S}^{(\mathrm{t})}, \mathcal{V})}. \end{aligned}$$

We set $1-\pi^{(\mathrm{v},\mathrm{o})}$ and $\pi^{(\mathrm{v},\mathrm{e})}$ to $\pi_0^{(\mathrm{v},\mathrm{p})}$ and $\pi_1^{(\mathrm{v},\mathrm{p})}$, respectively, so that the acceptance ratio can be calculated by

$$\begin{aligned} r^* = & \frac{p(V_{m,l}\neq v|V_{m,l-1})p(V_{m,r+1}|V_{m,r}\neq v)}{p(V_{m,l}=v|V_{m,l-1})p(V_{m,r+1}|V_{m,r}=v)} \\ & \times \frac{p(V_{m,n}\neq v|V_{m,n-1}=v)^{[l\neq 1\wedge V_{l-1}\neq v]+[r\neq N\wedge V_{r+1}\neq v]}}{p(V_{m,n}=v|V_{m,n-1}\neq v)^{[l\neq 1\wedge V_{l-1}=v]+[r\neq N\wedge V_{r+1}=v]}} \\ & \times \frac{p(\mathcal{X}^{(\mathrm{n})}|\mathcal{S}^{(\mathrm{n})}, \mathcal{V}^*)p(\mathcal{X}^{(\mathrm{t})}|\mathcal{S}^{(\mathrm{n})}, \mathcal{S}^{(\mathrm{t})}, \mathcal{V}^*)}{p(\mathcal{X}^{(\mathrm{n})}|\mathcal{S}^{(\mathrm{n})}, \mathcal{V})p(\mathcal{X}^{(\mathrm{t})}|\mathcal{S}^{(\mathrm{n})}, \mathcal{S}^{(\mathrm{t})}, \mathcal{V})}. \end{aligned}$$

2.5. *Efficient sampling from multimodal posteriors*

In addition to the standard sampling approaches mentioned earlier, we applied some additional elaborate sampling schemes to prevent the parameters from becoming stuck in a local optimum. One such scheme swaps parts of the nondecoy and decoy HLA sequences. First, a nondecoy index $m \in \{1, 2\}$, decoy index $m' \in \{3, \cdots, \nu^{(\mathrm{d})}+2\}$, and interval i such that $\forall n \in i; V_{m',n} = 1$ are sampled uniformly. Next, $S_{m,n}^{(\mathrm{n})}$ and $S_{m',n}^{(\mathrm{n})}$, and $S_{m,n}^{(\mathrm{t})}$ and $S_{m',n}^{(\mathrm{t})}$ are swapped for all $n \in i$. Finally, R_m^*, $R_{m'}^*$, $\mathcal{I}^{(\mathrm{n})*}$, and $\mathcal{I}^{(\mathrm{t})*}$ are sampled using Gibbs sampling given $\mathcal{S}^{(\mathrm{n})*}$ and $\mathcal{S}^{(\mathrm{t})*}$, which are the normal and tumor HLA sequences after swapping. Consequently, the acceptance ratio r^* is given by

$$r^* = \frac{p(\mathcal{X}^{(\mathrm{n})}|\mathcal{S}^{(\mathrm{n})*}, \mathcal{V})p(\mathcal{X}^{(\mathrm{t})}|\mathcal{S}^{(\mathrm{n})*}, \mathcal{S}^{(\mathrm{t})*}, \mathcal{V})p(\mathcal{S}^{(\mathrm{n})*})}{p(\mathcal{X}^{(\mathrm{n})}|\mathcal{S}^{(\mathrm{n})}, \mathcal{V})p(\mathcal{X}^{(\mathrm{t})}|\mathcal{S}^{(\mathrm{n})}, \mathcal{S}^{(\mathrm{t})}, \mathcal{V})p(\mathcal{S}^{(\mathrm{n})})}.$$

This sampling method helps to determine which HLA sequences should be decoys.

Another scheme involves sampling an HLA type and matched normal-tumor HLA sequences simultaneously. For all $m \in \{1, \ldots, \nu^{(\mathrm{d})}+2\}$, $S_m^{(\mathrm{n},\mathrm{N})}$ and $S_m^{(\mathrm{t},\mathrm{N})}$ are defined by

$$\begin{aligned} S_{m,n}^{(\mathrm{n},\mathrm{N})} &= \begin{cases} S_{m,n}^{(\mathrm{n})} & (\mathrm{if} D_{m,n} > 0) \\ \mathrm{N} & (\mathrm{if} D_{m,n} = 0) \end{cases}, \\ S_{m,n}^{(\mathrm{t},\mathrm{N})} &= \begin{cases} S_{m,n}^{(\mathrm{t})} & (\mathrm{if} D_{m,n} > 0) \\ \mathrm{N} & (\mathrm{if} D_{m,n} = 0) \end{cases}, \\ D_{m,n} &= D_{m,n}^{(\mathrm{n})} + D_{m,n}^{(\mathrm{t})} + D_{m+\nu^{(\mathrm{d})}+2,n}^{(\mathrm{t})}, \\ D_{m,n}^{(\mathrm{n})} &= |\{(i,j)|I_i^{(\mathrm{n})} = m, n \in r_{i,j}\}|, \\ D_{m,n}^{(\mathrm{t})} &= |\{(i,j)|I_i^{(\mathrm{t})} = m, n \in r_{i,j}\}| \end{aligned}$$

In other words, $S_m^{(\mathrm{n},\mathrm{N})}$ and $S_m^{(\mathrm{t},\mathrm{N})}$ are basically the same as $S_m^{(\mathrm{n})}$ and $S_m^{(\mathrm{t})}$, and bases not covered by any read are replaced with Ns. Next, R_m^* is sampled given $S_m^{(\mathrm{n},\mathrm{N})}$, $S_m^{(\mathrm{n})*}$ is sampled given R_m^* and $S_m^{(\mathrm{t},\mathrm{N})}$, and $S_m^{(\mathrm{t})*}$ is sampled given $S_m^{(\mathrm{n})*}$ in order. Then, the acceptance ratio r^* is given by

$$r^* = \frac{p(S_m^{(\mathrm{n},\mathrm{N})}|R_m)}{p(S_m^{(\mathrm{n},\mathrm{N})}|R_m^*)} \frac{p(\mathcal{X}^{(\mathrm{t})}|\mathcal{S}^{(\mathrm{n})}, \mathcal{S}^{(\mathrm{t},\mathrm{N})}, \mathcal{I}^{(\mathrm{t})})}{p(\mathcal{X}^{(\mathrm{t})}|\mathcal{S}^{(\mathrm{n})*}, \mathcal{S}^{(\mathrm{t},\mathrm{N})}, \mathcal{I}^{(\mathrm{t})})} \times \frac{p(S_m^{(\mathrm{t},\mathrm{N})}|\mathcal{S}_m^{(\mathrm{n})})}{p(S_m^{(\mathrm{t},\mathrm{N})}|\mathcal{S}_m^{(\mathrm{n})*})} \frac{p(S_m^{(\mathrm{t},\mathrm{N})}, \mathcal{X}^{(\mathrm{n})}, \mathcal{X}^{(\mathrm{t})}|R_m{}^*, \mathcal{I}^{(\mathrm{n})}, \mathcal{I}^{(\mathrm{t})})}{p(S_m^{(\mathrm{t},\mathrm{N})}, \mathcal{X}^{(\mathrm{n})}, \mathcal{X}^{(\mathrm{t})}|R_m, \mathcal{I}^{(\mathrm{n})}, \mathcal{I}^{(\mathrm{t})})} \frac{p(\mathcal{X}^{(\mathrm{t})}|\mathcal{S}^{(\mathrm{n})*}, \mathcal{S}_{-m}^{(\mathrm{t})}, \mathcal{I}^{(\mathrm{t})})}{p(\mathcal{X}^{(\mathrm{t})}|\mathcal{S}^{(\mathrm{n})}, \mathcal{S}_{-m}^{(\mathrm{t})}, \mathcal{I}^{(\mathrm{t})})}.$$

This sampling functions in a similar way to blocked Gibbs sampling of R_m, $S_m^{(\mathrm{n})}$, and $S_m^{(\mathrm{t})}$. This blocked Gibbs sampling requires substantial computation time because $S_m^{(\mathrm{n})}$ and $S_m^{(\mathrm{t})}$ must be integrated out for each HLA type. By contrast, our scheme requires much less time because $S_m^{(\mathrm{n})}$ and $S_m^{(\mathrm{t})}$ are integrated out only for R_m and R_m^*.

Other strategies were further used to obtain better parameters. First, reference sequences are periodically copied to HLA sequences. Second, sequence reads are assigned to decoy sequences if there are mismatches between the sequence reads and the reference sequences. These approaches help to reduce the incidence of false-positive mutations and retain only the mutations that seem true. The multistart method is also used to obtain better initial parameters. Moreover, parallel tempering is used to move parameters from mode to mode.

2.6. Human leukocyte antigen analysis from sampled parameters

HLA analysis is conducted based on the sampled parameters. HLA genotyping is performed by counting the number of sampled HLA types, and germline or somatic mutations are identified by finding different bases between HLA types and normal HLA sequences, or between normal and tumor HLA sequences, respectively.

3. RESULTS

3.1. Human leukocyte antigen genotyping from whole-genome sequencing data

We first evaluated the accuracy of this method for HLA genotyping from a WGS data set. For comparison, we applied ALPHLARD-NT, ALPHLARD (Hayashi et al., 2018), and POLYSOLVER (Shukla et al., 2015) to WGS data of 25 colon cancer samples, which were used by Hayashi et al. (2018). The performance comparison is summarized in Table 1. Overall, ALPHLARD-NT outperformed POLYSOLVER at all resolutions for all HLA loci. ALPHLARD-NT also achieved slightly higher accuracy than ALPHLARD because ALPHLARD-NT can use information from both normal and tumor samples, whereas ALPHLARD can only use information from normal samples.

3.2. Detection of human leukocyte antigen mutations from whole-genome sequencing data

We also searched for HLA class I somatic mutations among the WGS data from the 25 colon cancer samples using ALPHLARD-NT, POLYSOLVER, and EBCall (Shiraishi et al., 2013), which is a standard mutation caller. ALPHLARD-NT called one substitution, two insertions, and two deletions, all of which were verified by the TruSight HLA Sequencing Panels (Weimer et al., 2016). All four indels called are known to lead to the loss of function of the HLA alleles, and might contribute to immune escape. However, POLYSOLVER and EBCall detected no and one mutation, respectively, which was likely due to the low coverage of the data set.

3.3. Detection of human leukocyte antigen mutations from whole-exome sequencing data

Next, we applied ALPHLARD-NT, POLYSOLVER, and EBCall to a WES data set of 343 colon adenocarcinoma cases from The Cancer Genome Atlas (TCGA). Figure 2 shows the Venn diagrams of the identified HLA class I somatic mutations with each method. This figure demonstrates the high sensitivity of ALPHLARD-NT (88 mutations) compared with POLYSOLVER (60 mutations) and

Table 1. Comparison of the Accuracy of Whole-Genome Sequencing-Based Human Leukocyte Antigen Genotyping with ALPHLARD-NT, ALPHLARD, and POLYSOLVER

	ALPHLARD-NT	*ALPHLARD*	*POLYSOLVER*
HLA-A			
First	**100% (50/50)**	**100% (50/50)**	**100% (50/50)**
Second	**100% (50/50)**	98.0% (49/50)	98.0% (49/50)
Third	**98.0% (49/50)**	**98.0% (49/50)**	90.0% (45/50)
HLA-B			
First	**100% (48/48)**	**100% (48/48)**	91.7% (44/48)
Second	**100% (48/48)**	**100% (48/48)**	85.4% (41/48)
Third	**97.9% (47/48)**	95.8% (46/48)	81.3% (39/48)
HLA-C			
First	**100% (50/50)**	**100% (50/50)**	**100% (50/50)**
Second	**100% (50/50)**	98.0% (49/50)	90.0% (45/50)
Third	**100% (50/50)**	98.0% (49/50)	86.0% (43/50)
HLA-DPA1			
First	**100% (24/24)**	**100% (24/24)**	N/A
Second	**100% (24/24)**	**100% (24/24)**	N/A
Third	**100% (24/24)**	**100% (24/24)**	N/A
HLA-DPB1			
First	**100% (22/22)**	**100% (22/22)**	N/A
Second	**100% (22/22)**	**100% (22/22)**	N/A
Third	**100% (22/22)**	**100% (22/22)**	N/A
HLA-DQA1			
First	**100% (24/24)**	**100% (24/24)**	N/A
Second	**95.8% (23/24)**	**95.8% (23/24)**	N/A
Third	**95.8% (23/24)**	**95.8% (23/24)**	N/A
HLA-DQB1			
First	**100% (18/18)**	**100% (18/18)**	N/A
Second	**94.4% (17/18)**	**94.4% (17/18)**	N/A
Third	**94.4% (17/18)**	**94.4% (17/18)**	N/A
HLA-DRB1			
First	**100% (24/24)**	**100% (24/24)**	N/A
Second	**100% (24/24)**	**100% (24/24)**	N/A
Third	**100% (24/24)**	**100% (24/24)**	N/A
Total			
First	**100% (260/260)**	**100% (260/260)**	97.3% (144/148)
Second	**99.2% (258/260)**	98.5% (256/260)	91.2% (135/148)
Third	**98.5% (256/260)**	97.7% (254/260)	85.8% (127/148)

N/A indicates that the method does not support the HLA locus.
HLA, human leukocyte antigen.
Bold values indicate that the method achieved the highest accuracy for the HLA locus at the resolution.

EBCall (80 mutations), which is especially remarkable for insertions. ALPHLARD-NT detected seven insertions at the beginning of exon 4 of HLA class I genes, which is a known hotspot of indels (Mizuno et al., 2018), whereas POLYSOLVER and EBCall identified no and three insertions at this hotspot, respectively. ALPHLARD-NT also identified 12 deletions at the same position. These recurrent frameshift indels seemed to be positively selected for immune escape caused by loss of function of the HLA alleles.

In addition, ALPHLARD-NT detected a novel HLA-B allele whose exon sequence is the same as HLA-B*35:08:01 except that the 25th base is C rather than G, which changes the 9th amino acid from V to L. The protein produced by the new allele is also novel and not registered in the IPD-IMGT/HLA Database, indicating that the allele defines a new HLA type name at the second field.

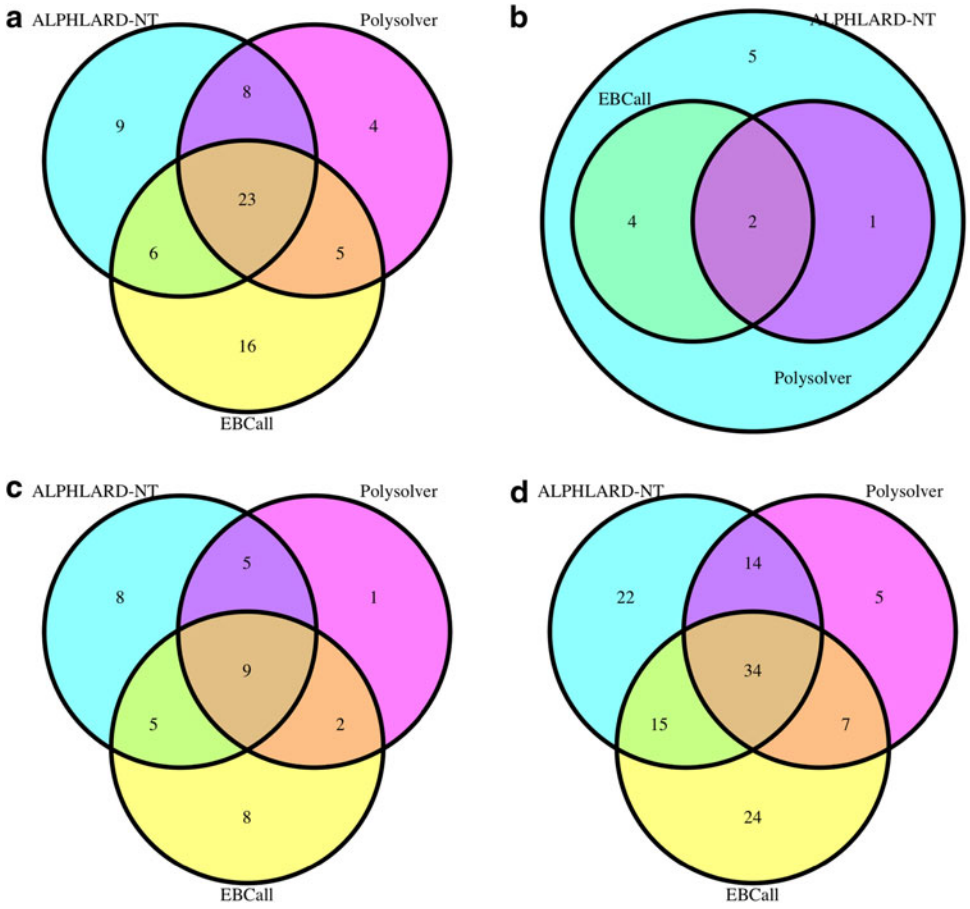


FIG. 2. Venn diagrams of the number of HLA somatic mutations identified by ALPHLARD-NT, POLYSOLVER, and EBCall for **(a)** substitutions, **(b)** insertions, **(c)** deletions, and **(d)** all mutations. HLA, human leukocyte antigen.

Algorithm 1 Generate a candidate parameter V^* using the Wolff algorithm

Input:

V: the current parameter
N: the length of V
$\pi_0^{(\mathrm{v},\mathrm{p})}$: probability for 0-cluster extension
$\pi_1^{(\mathrm{v},\mathrm{p})}$: probability for 1-cluster extension

Output:

V^*: candidate parameter

function WOLFF (V, $\pi_0^{(\mathrm{v},\mathrm{p})}$, $\pi_1^{(\mathrm{v},\mathrm{p})}$)
  Sample a position p uniformly
  $v \leftarrow V_p$
  $b \leftarrow p$
  while $b > 1$ **and** $V_{b-1} = v$ **do**
    break with probability $1 - \pi_v^{(\mathrm{v},\mathrm{p})}$
    $b \leftarrow b - 1$
  end while
  $e \leftarrow p$
  while $e < N$ **and** $V_{e+1} = v$ **do**
    break with probability $1 - \pi_v^{(\mathrm{v},\mathrm{p})}$
    $e \leftarrow e + 1$
  end while
  $V^* \leftarrow V$
  for $n \leftarrow b$ to e **do**
    $V_n^* \leftarrow 1 - v$
  end for
  return V^*
end function

4. CONCLUSION

In this article, we have presented a new Bayesian method, ALPHLARD-NT, which identifies HLA germline and somatic mutations as well as HLA genotypes. Comparison of the performance of ALPHLARD-NT clearly demonstrated its higher accuracy than existing methods for WGS-based HLA genotyping. ALPHLARD-NT also detected HLA somatic mutations from both WES and WGS data. In general, HLA mutation calling is difficult mainly due to the similarity of HLA genes and pseudogenes. We dealt with this problem by applying sophisticated filtering criteria and using decoy-related parameters that reduced the influence of misclassified reads at the filtering step. Although these approaches work well for HLA class I mutation calling, identification of HLA class II mutations remains a challenge, since databases tend to be relatively incomplete for identifying class II genes and pseudogenes compared with class I genes.

With the continuous accumulation of large amounts of WES and WGS data, HLA mutation calling from these data sets is a fundamental step in cancer immunogenomics. Thus, we expect that our method will be an essential tool for comprehensive analyses of HLA genes from WES and WGS data.

ACKNOWLEDGMENT

The super-computing resource was provided by Human Genome Center, the Institute of Medical Science, the University of Tokyo.

AUTHOR DISCLOSURE STATEMENT

The authors declare there are no competing financial interests.

REFERENCES

Bai, Y., Ni, M., Cooper, B., et al. 2014. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. *BMC Genomics*. 15, 325.

Boegel, S., Löwer, M., Schäfer, M., et al. 2012. HLA typing from RNA-Seq sequence reads. *Genome Med*. 4, 102.

Dilthey, A.T., Gourraud, P.-A., Mentzer, A.J., et al. 2016. High-accuracy HLA type inference from whole-genome sequencing data using population reference graphs. *PLoS Comput. Biol*. 12, e1005151.

Giannakis, M., Mu, X.J., Shukla, S.A., et al. 2016. Genomic correlates of immune-cell infiltrates in colorectal carcinoma. *Cell Rep*. 15, 857–865.

González-Galarza, F.F., Takeshita, L.Y., Santos, E.J., et al. 2015. Allele frequency net 2015 update: New features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res*. 43, D784–D788.

Grivennikov, S.I., Greten, F.R., and Karin, M. 2010. Immunity, inflammation, and cancer. *Cell* 140, 883–899.

Hayashi, S., Yamaguchi, R., Mizuno, S., et al. 2018. ALPHLARD: A Bayesian method for analyzing HLA genes from whole genome sequence data. *BMC Genomics* 19, 790.

Kim, H.J., and Pourmand, N. 2013. HLA haplotyping from RNA-seq data using hierarchical read weighting. *PLoS One* 8, e67885.

Kreiter, S., Vormehr, M., Van de Roemer, N., et al. 2015. Mutant MHC class II epitopes drive therapeutic immune responses to cancer. *Nature* 520, 692.

Lee, H., and Kingsford, C. 2018. Kourami: Graph-guided assembly for novel human leukocyte antigen allele discovery. *Genome Biol*. 19, 16.

Liu, C., Yang, X., Duffy, B., et al. 2013. ATHLATES: Accurate typing of human leukocyte antigen through exome sequencing. *Nucleic Acids Res*. 41, e142.

Marty, R., Kaabinejadian, S., Rossell, D., et al. 2017. MHC-I genotype restricts the oncogenic mutational landscape. *Cell* 171, 1272–1283.

McGranahan, N., Rosenthal, R., Hiley, C.T., et al. 2017. Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell* 171, 1259–1271.

Mizuno, S., Yamaguchi, R., Hasegawa, T., et al. 2018. Immuno-genomic PanCancer landscape reveals diverse immune escape mechanisms and immuno-editing histories. *bioRxiv*, 285338.

Nariai, N., Kojima, K., Saito, S., et al. 2015. HLA-VBSeq: Accurate HLA typing at full resolution from whole-genome sequencing data. *BMC Genomics* 16, S7.

Olerup, O., and Zetterquist, H. 1992. HLA-DR typing by PCR amplification with sequence-specific primers (PCR-SSP) in 2 hours: An alternative to serological DR typing in clinical practice including donor-recipient matching in cadaveric transplantation. *Tissue Antigens* 39, 225–235.

Robinson, J., Halliwell, J.A., Hayhurst, J.D., et al. 2015. The IPD and IMGT/HLA database: Allele variant databases. *Nucleic Acids Res*. 43, D423–D431.

Rooney, M.S., Shukla, S.A., Wu, C.J., et al. 2015. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* 160, 48–61.

Saiki, R.K., Bugawan, T.L., Horn, G.T., et al. 1986. Analysis of enzymatically amplified β-globin and HLA-DQα DNA with allele-specific oligonucleotide probes. *Nature* 324, 163.

Santamaria, P., Boyce-Jacino, M.T., Lindstrom, A.L., et al. 1992. HLA class II ''typing'': Direct sequencing of DRB, DQB, and DQA genes. *Hum. Immunol*. 33, 69–81.

Schreiber, R.D., Old, L.J., and Smyth, M.J. 2011. Cancer immunoediting: Integrating immunity's roles in cancer suppression and promotion. *Science* 331, 1565–1570.

Shiraishi, Y., Sato, Y., Chiba, K., et al. 2013. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res*. 41, e89.

Shukla, S.A., Rooney, M.S., Rajasagi, M., et al. 2015. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol*. 33, 1152–1158.

Szolek, A., Schubert, B., Mohr, C., et al. 2014. OptiType: Precision HLA typing from next-generation sequencing data. *Bioinformatics* 30, 3310–3316.

Testoni, M., Zucca, E., Young, K., et al. 2015. Genetic lesions in diffuse large B-cell lymphomas. *Ann. Oncol*. 26, 1069–1080.

The Cancer Genome Atlas Network. 2015. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* 517, 576.

The Cancer Genome Atlas Research Network. 2014. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513, 202.

Warren, R.L., Choe, G., Freeman, D.J., et al. 2012. Derivation of HLA types from shotgun sequence datasets. *Genome Med*. 4, 95.

Weimer, E.T., Montgomery, M., Petraroia, R., et al. 2016. Performance characteristics and validation of next-generation sequencing for human leucocyte antigen typing. *J. Mol. Diagn*. 18, 668–675.

Wolff, U. 1989. Collective Monte Carlo updating for spin systems. *Phys. Rev. Lett*. 62, 361.

Xie, C., Yeo, Z.X., Wong, M., et al. 2017. Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. *Proc. Natl Acad. Sci. U. S. A*. 114, 8059–8064.

Address correspondence to:
Professor Seiya Imoto
Health Intelligence Center
The Institute of Medical Science
The University of Tokyo
4-6-1 Shirokanedai
Minato-ku
Tokyo 108-8639
Japan

E-mail: imoto@ims.u-tokyo.ac.jp