Bayesian Statistical Methods for Comprehensive Analysis of HLA Genes from Whole Genome Sequence Data
（全ゲノムシークエンスデータを用いたHLA遺伝子の網羅的解析に対するベイズ統計的手法）

by

Shuto Hayashi
林 周斗

A Doctor Thesis
博士論文

Submitted to
the Graduate School of the University of Tokyo
on December 7, 2018
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Information Science and Technology
in Computer Science

Thesis Supervisor: Satoru Miyano　宮野 悟
Professor of Computer Science

ABSTRACT

Human leukocyte antigen (HLA) genes are essential components of the immune system, which facilitate the elimination of virus-infected cells. HLA genes must be highly diverse and have a lot of single nucleotide polymorphisms (SNPs) in the human genome to protect against various kinds of viruses. These polymorphism patterns in DNA sequences define HLA types, or alleles, in each HLA gene. Different HLA types show different immune responses because the binding affinity of an HLA molecule and a peptide differs depending on the HLA type, resulting in high individual variation in immune responses including disease susceptibility. Therefore, HLA genotyping, in which the specific pair of HLA types is identified for each HLA locus, is essential to understand the immune system. In addition, researchers have focused on the interaction between cancer and the immune system because tumor cells could be also killed by the immune system. Recent studies have shown that somatic mutations in HLA genes tend to accumulate in specific cancer types. Since these HLA somatic mutations have the potential to change immune responses, HLA somatic mutation calling as well as HLA genotyping can further help to understand the link between cancer and immunity.

Recently, HLA genotyping from next-generation sequencing (NGS) data has attracted attention as NGS technologies have become an essential tool to analyze DNA or RNA because they have achieved high throughput sequence data at low costs. There are several types of NGS data such as whole exome sequence (WES) data, whole genome sequence (WGS) data, and RNA sequence (RNA-seq) data. A lot of NGS data have been generated, stored, and shared, and hence it is important to take advantage of such a large amount of NGS data. However, HLA genotyping from NGS data is difficult due to some reasons. First, the number of possible combinations of HLA types is enormous. Therefore, it is impractical to obtain the best HLA genotype of a sample by checking all of the possible HLA genotypes. Second, there are several dozens of HLA genes and HLA pseudogenes in total, and they have quite similar DNA sequences to each other. Hence, it is difficult to judge which HLA gene or HLA pseudogene produced each sequence read.

A number of methods have been developed to tackle these problems and perform HLA genotyping from NGS data. Some of these methods have achieved sufficiently high accuracy for HLA genotyping from WES and RNA-seq data. However, it has been reported that these methods cannot accurately determine HLA genotypes from WGS data. Besides, no methods have achieved accurate HLA mutation calling from WGS data. In this thesis, we tackle these problems.

First, we introduce a method to extract and classify sequence reads from HLA genes, which is necessary for subsequent HLA analysis. The extraction and classification of HLA reads are basically difficult because of the high similarity in HLA genes and HLA pseudogenes. We deal with this problem using an original alignment scoring that reduces misclassification of sequence reads by considering not only the number of mismatches but also base qualities at the mismatch positions.

Second, we propose a new Bayesian method, called ALPHLARD, that accurately determines HLA genotypes from WGS data as well as WES data. ALPHLARD conducts HLA genotyping for each HLA locus independently by using reads that were classified into the HLA locus. The model incorporates the parameters for not only HLA types but also HLA sequences of the sample, which make it possible to detect HLA germline mutations and identify new HLA types that are not registered in the HLA type database by checking differences between the HLA types and the HLA sequences. Moreover, we add the parameters of decoy HLA types and decoy HLA sequences to the model, which reduce the influence of misclassified sequence reads that are really produced by other HLA genes and HLA pseudogenes than the HLA gene of interest. ALPHLARD estimates the HLA genotype and the HLA germline mutations by calculating the posterior distribution using the Markov chain Monte Carlo (MCMC) method. To accelerate the MCMC convergence, we introduce several proposal distributions that enable parameters to jump from mode to mode of the posterior distribution. We compared ALPHLARD with other existing methods using WES data and WGS data, and confirmed that ALPHLARD outperformed the other methods in the accuracy of HLA genotyping.

Finally, we propose a method, called ALPHLARD-NT, to conduct HLA somatic mutation calling as well as HLA genotyping and HLA germline mutation calling from normal and tumor sequence data of cancer patients. ALPHLARD-NT performs HLA genotyping, HLA germline mutation calling, and HLA somatic mutation calling through simultaneous analysis of both normal and tumor sequence data, although existing methods conduct these procedures separately. The statistical model of ALPHLARD-NT is obtained by extending ALPHLARD to include additional parameters for tumor sequence data. We also add parameters that control the ratio of sequence reads that are produced by each HLA sequence. As with ALPHLARD, ALPHLARD-NT also uses MCMC to estimate the posterior distribution of the parameters. We compared ALPHLARD-NT with existing methods using WES data and WGS data, and validated that ALPHLARD-NT could sensitively identify HLA somatic mutations even from WGS data.

論文要旨

ヒト白血球抗原（HLA）遺伝子は免疫系の重要な構成要素であり、ウイルス感染細胞を排除する手助けをする。種々様々なウイルスから身を守るため、HLA遺伝子はヒトゲノムの中でも非常に多様な領域であり、多くの一塩基多型（SNP）を持つ。DNA配列におけるこれらの多型のパターンによって、それぞれのHLA遺伝子においてHLA型が定義される。HLA分子とペプチドの結合親和性はHLA型ごとに異なるため、HLA型ごとに免疫反応は異なる。その結果、病気への罹りやすさをなどの免疫反応は個々人ごとに大きく異なる。それゆえ、HLA遺伝子型決定、すなわちそれぞれのHLA遺伝子に対して一組のHLA型を同定することは免疫系を理解する上で重要である。また、腫瘍細胞もまた免疫系に排除されうるため、研究者たちはがんと免疫系の相互作用に注目している。近年の研究により、HLA遺伝子における体細胞変異は特定のがん種に蓄積する傾向があることが分かった。これらのHLA体細胞変異は免疫反応を変えうるため、HLA遺伝子型決定だけでなくHLA体細胞変異同定もまたがんと免疫の繋がりを理解する手助けとなりうる。

近年、次世代シークエンス（NGS）の技術により安価に大量のデータが手に入るようになり、NGS技術は重要なツールになってきたため、NGSデータからHLA遺伝子型決定を行うことが注目されている。NGSデータには全エクソームシークエンス（WES）データ、全ゲノムシークエンス（WGS）データ、RNAシークエンス（RNA-seq）データのようないくつかの種類がある。多くのNGSデータが生成、保存、共有されてきているため、そのような大量のNGSデータを活用することが重要である。しかしながら、シークエンスデータを用いてHLA遺伝子型決定を行うことはいくつかの理由から困難である。まず、可能なHLA型の組み合わせの数は膨大であることが挙げられる。そのため、全てのHLA遺伝子型を探索することにより最適なHLA遺伝子型を得るのは現実的でない。また、HLA遺伝子やHLA偽遺伝子は合わせて数十あり、それらはお互いに似たDNA配列を持つことが挙げられる。それゆえ、それぞれのシークエンスリードがどのHLA遺伝子あるいはHLA偽遺伝子によって生成されたのか判定するのが困難である。

これらの問題に取り組み、シークエンスデータを用いてHLA遺伝子型決定を行うため、数多くの手法が開発されてきた。これらの手法の中にはWESデータやRNA-seqデータから十分高精度にHLA遺伝子型決定できるものがある。しかしながら、これらの手法はWGSデータから高精度にHLA遺伝子型決定を行えないことが報告されている。加えて、どの手法もWGSデータから精度の高いHLA変異同定を行えてはいない。この学位論文ではこれらの問題に取り組む。

まず、我々はHLA遺伝子から生成されたシークエンスリードを抽出し、分類する手法を紹介する。これは後のHLA解析に必要となることである。HLA遺伝子とHLA偽遺伝子はお互いに非常に似ているため、HLAリードの抽出と分類は基本的に難しい。我々はミスマッチの数だけでなくミスマッチが起こった箇所におけるベースクオリティも考慮することにより、シークエンスリードの誤分類を減らす独自のアライメントスコアリングを用いてこの問題に対処する。

次に、我々はWESデータだけでなくWESデータからも正確にHLA遺伝子型を決定する新たなベイズ的手法、ALPHLARDを提案する。ALPHLARDは各HLA遺伝子に分類されたリードを用いることでそれぞれのHLA遺伝子ごとに独立にHLA遺伝子型決定を行う。ALPHLARDのモデルはHLA型だけでなくそのサンプルのHLA配列に対するパラ

メータも含んでいる。これにより、HLA 型と HLA 配列の違いを確認することで、HLA 生殖細胞変異を検出し、HLA 型データベースに登録されていない新たな HLA 型を同定することができる。さらに、我々はモデルにデコイ HLA 型とデコイ HLA 配列のパラメータを追加する。これにより、実際には注目している HLA 遺伝子以外の HLA 遺伝子や HLA 偽遺伝子から生成されたリードの誤分類の影響を減らすことができる。ALPHLARD ではマルコフ連鎖モンテカルロ法（MCMC）を用いて事後分布を計算することにより、HLA 遺伝子型と HLA 生殖細胞変異を推定する。MCMC の収束を加速するため、我々はパラメータが事後分布のモードからモードへ移動できるような提案分布をいくつか導入する。我々は WES データと WGS データを用いて ALPHLARD と既存手法を比較し、ALPHLARD が既存手法より精度の高い HLA 遺伝子型決定を行えることを確認した。

最後に、我々はがん患者の正常細胞と腫瘍細胞のシークエンスデータから、HLA 遺伝子型決定や HLA 生殖細胞変異同定だけでなく HLA 体細胞変異同定も行う手法、ALPHLARD-NT を提案する。既存手法では正常細胞と腫瘍細胞シークエンスデータを別々に解析するが、ALPHLARD-NT はこれらを両方同時に解析することで HLA 遺伝子型決定、HLA 生殖細胞変異同定、HLA 体細胞変異同定を行う。ALPHLARD-NT の統計モデルは腫瘍細胞のシークエンスデータに対する追加パラメータを含むように ALPHLARD を拡張することで得られる。さらに我々はそれぞれの HLA 配列から生成されたシークエンスリードの比を制御するパラメーラも追加する。ALPHLARD と同様に、ALPHLARD-NT でもパラメータの事後分布の推定に MCMC を用いる。我々は WES データと WGS データを用いて ALPHLARD-NT と既存手法を比較し、ALPHLARD-NT が WGS データからでも HLA 体細胞変異を高感度で同定することができることを確認した。

Acknowledgements

First of all, I would like to thank my supervisor, Professor Satoru Miyano, for educating me for more than five years. He has provided me a good research environment where I have been able to focus on my research. With his support, I could have accomplished this work.

I am also grateful to Professor Seiya Imoto and Associate Professor Rui Yamaguchi for giving me helpful advice on my work. They have supported my research activities and have also cultivated my academic writing and presentation skills.

In addition, I appreciate valuable feedbacks from my collaborators, Professor Shinichi Mizuno, Dr. Hidewaki Nakagawa, Dr. Takanori Hasegawa, Mitsuhiro Komura, and Takuya Moriyama. Dr. Nakagawa also provided me with precious sequence data, which are used in this thesis.

Moreover, I would like to all members of Laboratory of DNA Information Analysis and Laboratory of Sequence Analysis in Human Genome Center, and Division of Health Medical Data Science and Division of Health Medical Computational Science in Health Intelligence Center. Especially, I am thankful to Dr. Satoshi Ito, Dr. Masanori Kakuta, Dr. Zhang Yao-zhong, Dr. Georg Tremmel, Ayako Tomiyasu, Asako Suzuki, Dr. Taku Onodera, Dr. Rika Kasajima, Eigo Shimizu and Dr. Atsushi Niida for giving me helpful comments and supports. I am also grateful to Professor Teppei Shimamura and Dr. Yuichi Shiraishi for meaningful discussions. Lastly, I wish to thank my parents for supporting and encouraging me.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Overview

Human leukocyte antigen (HLA) genes are essential components of the immune system, which facilitate elimination of virus-infected cells. HLA genes are mainly classified into two categories: HLA class I genes including the HLA-A, HLA-B, and HLA-C genes, and HLA class II genes such as HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, and HLA-DRB1 genes. HLA class I genes are expressed in almost all cells, and the function is binding to peptides with high affinity and bringing them to the cell surface. If a cell is infected by a virus, HLA class I molecules could deliver viral peptides to the cell surface, and the viral-peptide/HLA complexes would be recognized by killer T cells. As a result, the virus-infected cells are attacked and destroyed by activated T cells. On the other hand, although self peptides could be also presented to T cells, the self-peptide/HLA complexes are not be recognized by T cells because T cells are trained not to recognize self-peptide/HLA complexes in thymus, which means that non-infected cells are not attacked by T cells. Thus, only virus-infected cells are selectively eliminated by the immune system. HLA class II genes are expressed only in specific types of cells, called antigen presenting cells, which consist of dendritic cells, monocytes, macrophages, and B cells. If an APC internalizes a virus, HLA class II molecules could display viral peptides on the cell surface, and the viral-peptide/HLA complexes would be recognized by helper T cells. Then, helper T cells activate immune cells, including killer T cells and B cells, and the activated immune cells work to remove viruses in the body.

Since an HLA molecule can bind to some viral peptides but cannot bind to others, HLA genes must be highly diverse and have a lot of single nucleotide polymorphisms (SNPs) in the human genome to protect against various kinds of viruses (Figure 1.1). These polymorphism patterns in DNA sequences define HLA types, or alleles, in each HLA gene. For example, in the case of HLA class I genes, the HLA-A, HLA-B, and HLA-C genes have 4,638, 5,590, and 4,374 types registered in the IPD-IMGT/HLA Database (Release 3.34.0) [69], which is one of the most popular databases of HLA types. On the other hand, HLA class II genes are less polymorphic than HLA class I genes, and the HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, and HLA-DRB1 genes have 73, 1,097, 100, 1,316, 2,300 types registered in the IPD-IMGT/HLA Database. Different HLA types show different immune responses because the binding affinity of an HLA molecule and a peptide differs depending on the HLA type, resulting in high individual variation in immune responses including disease susceptibility [75, 65, 24, 59, 8, 17, 33, 22, 7, 52, 63, 67]. Since the human genome is diploid, humans have two HLA types in each HLA locus, and identifying the specific pair

HLA type	DNA sequence
HLA-A*02:01:01:01	...ACTCACCGAGTGGACCTGGGGACCCTGCGCGGC...
HLA-A*03:01:01:01	...ACTGACCGAGTGGACCTGGGGACCCTGCGCGGC...
HLA-A*11:01:01:01	...ACTGACCGAGTGGACCTGGGGACCCTGCGCGGC...
HLA-A*24:02:01:01	...ACTGACCGAGAGAACCTGCGGATCGCGCTCCGC...
HLA-A*26:01:01:01	...ACTGACCGAGCGAACCTGGGGACCCTGCGCGGC...

Figure 1.1: An example of single nucleotide polymorphisms (SNPs) in the HLA-A gene.

of HLA types for each HLA locus is called HLA genotyping. Figure 1.2 shows an example of the result of HLA genotyping. HLA genotyping is essential to not only research on diseases mentioned above but also organ transplantation because HLA matching between the donor and the recipient would reduce the influence of rejection caused by attacks from the donor's T cells to the recipient's organ.

In addition, researchers have focused on the interaction between cancer and the immune system, and immune therapies for cancer [25, 76, 70, 40, 55] because tumor cells could be also killed by T cells through the same mechanism as elimination of virus-infected cells, in which mutated peptides are used instead of viral peptides. However, some tumor cells can evade the elimination by suppressing the ability of the immune system [68, 42, 16], which is called tumor immune escape. Recent studies have shown that somatic mutations in HLA genes tend to accumulate in specific cancer types [84, 78, 83, 82, 20, 56]. These HLA somatic mutations have potential to change immune responses, including tumor immune escape. Especially, insertions, deletions, and nonsense mutations, which are mutations that change an amino acid into a stop codon, are considered to be crucial because they generally cause the loss of function of HLA genes. Hence, HLA

HLA locus	HLA type 1	HLA type 2
HLA-A	HLA-A*26:03:01	HLA-A*31:01:02:01
HLA-B	HLA-B*35:01:01:01	HLA-B*35:01:01:01
HLA-C	HLA-C*03:03:01:01	HLA-C*03:03:01:01
HLA-DPA1	HLA-DPA1*01:03:01:01	HLA-DPA1*02:02:02:01
HLA-DPB1	HLA-DPB1*02:02:01:01	HLA-DPB1*16:01:01
HLA-DQA1	HLA-DQA1*01:02:01:01	HLA-DQA1*03:01:01
HLA-DQB1	HLA-DQB1*03:02:01:01	HLA-DQB1*06:02:01:01
HLA-DRB1	HLA-DRB1*04:03:01:01	HLA-DRB1*15:01:01:01

Figure 1.2: An example of the result of HLA genotyping.

somatic mutation calling as well as 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 [72], sequence-specific primers [66], and sequence-based typing [73]. However, these methods are time-consuming and labor-intensive, and can only provide information on targeted HLA genes. Also, the methods frequently cannot determine HLA genotypes uniquely because they do not use phase information on whether or not two SNPs are located on the same chromosome.

Recently, HLA genotyping from next-generation sequencing (NGS) data has attracted attention as NGS technologies have become an essential tool to analyze DNA or RNA because they have achieved high throughput sequence data at low costs. NGS data consists of sequenced DNA or RNA fragments of a sample, which are called sequence reads, and the number of sequenced base pairs of one sample reaches millions to trillions. There are several types of NGS data such as whole exome sequence (WES) data, whole genome sequence (WGS) data, and RNA sequence (RNA-seq) data. Each type of NGS data has its own characteristics. WES focuses only on exons, or protein coding regions, which account for 1-2% of the human genome. On the other hand, WGS sequences the entire genome including non-coding regions such as introns and intergenic regions. In the case of WES, exonic regions can be sequenced more repeatedly at lower costs, so that the influence of sequencing errors is relatively small. In other words, WES data provides information only on limited regions but generally enable more accurate analysis than WGS, which means that accurate genome analysis from WGS data requires sophisticated methods. RNA-seq captures all transcripts including not only messenger RNAs, which could be translated into proteins, but also non-coding RNAs. As with WES data, RNA-seq captures only expressed regions and can sequence the regions repeatedly at low costs. RNA-seq data is different from WES and WGS data in that the amount of sequence reads of each gene is different depending on the expression level of the gene.

In recent years, a lot of NGS data have been generated, stored, and shared. The Cancer Genome Atlas [86] is such a big cancer genome project and has NGS data of more than 11,000 patients across more than 30 cancer types. The International Cancer Genome Consortium [32] is also a big project and stores NGS data obtained from more than 15,000 patients across more than 20 cancer types. Therefore, it is important to take advantage of such a large amount of NGS data. However, HLA genotyping from NGS data is difficult due to some reasons. First, the number of possible combinations of HLA types is enormous. For example, the HLA-A gene has 10,757,841 possible HLA genotypes. Therefore, it is impractical to obtain the best HLA genotypes of a sample by checking all of the possible HLA genotypes. Second, there are several dozens of HLA genes and HLA pseudogenes in total, and they have quite similar DNA sequences to each other. Hence, it is difficult to judge which HLA gene or HLA pseudogene produced each sequence read.

A number of methods have been developed to tackle these problems and perform HLA genotyping from NGS data [15, 31, 88, 5, 51, 38, 3, 81, 62, 78, 11, 93, 45]. Some of these methods have achieved sufficiently high accuracy for HLA genotyping from WES and RNA-seq data. With the existing 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 [78, 45] are valuable. However, Bauer *et al.* has reported that these methods cannot reach 80% accuracy for HLA genotyping from WGS data [4]. Besides, no methods have achieved accurate HLA mutation calling from WGS data. These are because WGS data is relatively shallow and susceptible to the problem of high similarity among HLA genes and HLA pseudogenes, and it is difficult to distinguish true HLA mutations from false-positive mutations caused by similar HLA genes and HLA pseudogenes. Thus, HLA genotyping, and HLA germline and somatic mutation calling from WGS data remain a significant challenge, although this approach would provide more information of HLA loci than possible with WES and RNA-seq data, including details of the non-coding regions such as the introns and the untranslated regions.

1.2 Contributions of This Thesis

In this section, we briefly explain our contributions of this thesis.

1.2.1 Extraction, Classification, and Realignment of HLA Reads from Sequence Data

Analysis of HLA genes from NGS data generally begins with extraction of sequence reads from HLA genes. However, due to the high similarity in HLA genes and HLA pseudogenes, we also must judge which HLA gene or pseudogene produced each read. In Chapter 3, we introduce a method to extract, classify, and realign sequence reads from HLA genes. The method is based on an original alignment-based scoring that calculates how likely each read is to be produced by each HLA type.

1.2.2 Bayesian Approach for HLA Genotyping from Whole Genome Sequence Data

After realignment of HLA reads, we perform HLA genotyping using the realigned reads. In Chapter 4, we describe a Bayesian model, ALPHLARD, to accurately determines HLA genotypes. ALPHLARD also can identify HLA germline mutations by introducing parameters for HLA sequences as well as parameters for HLA types. In addition, ALPHLARD contains decoy parameters, which can reduce the influence of misclassified reads that are really produced by the other HLA genes and HLA pseudogenes than the HLA gene of interest. We use the Markov chain Monte Carlo (MCMC) method to sample parameters from the posterior distribution. We further introduce some efficient proposal distributions that enable parameters to jump from mode to mode. Experimental results show that ALPHLARD outperforms existing methods in HLA genotyping from both WES data and WGS data.

1.2.3 Bayesian Approach for HLA Somatic Mutation Calling from Whole Genome Sequence Data

In Chapter 5, we further introduce a Bayesian model, ALPHLARD-NT, that can identify HLA somatic mutations as well as HLA genotypes and HLA germline mutations. ALPHLARD-NT is constructed by extending ALPHLARD to contain some additional parameters: parameters for tumor sequence data and parameters to control the ratio of sequence reads that are produced by each HLA sequence. As with ALPHLARD, ALPHLARD-NT also uses MCMC to estimate the posterior distribution. Experimental results show that ALPHLARD-NT achieves higher accuracy than other methods in HLA genotyping from paired normal and tumor WGS data. They also demonstrate that ALPHLARD-NT can sensitively identify HLA somatic mutations compared with existing methods.

1.3 Organization of This Thesis

The rest of this thesis is organized as follows. In Chapter 2, we provide preliminary information. In Chapter 3, we introduce an alignment-based scoring method to extract, classify, and realign sequence reads from HLA genes. In Chapter 4, we present a Bayesian model to accurately determines HLA genotypes and HLA germline mutations from the realigned reads. In Chapter 5, we propose a Bayesian model to identify HLA somatic mutations as well as HLA genotypes and HLA germline mutations from the paired normal and tumor realigned reads. Finally, in Chapter 6, we conclude this thesis.

Chapter 2

Preliminaries

2.1 Nomenclature of HLA Types

HLA genes are located in a highly polymorphic region and have a lot of SNPs in the human genome. The polymorphism pattern defines an HLA type for each HLA locus, whose naming is managed by the WHO Nomenclature Committee for Factors of the HLA System [54]. In this section, we explain the nomenclature of HLA types. Figure 2.1 shows an example of an HLA type, HLA-A*02:01:01:02L. The name of each HLA type is separated by an asterisk. What is written to the left of the asterisk is the HLA gene. On the other hand, what is written to the right of the asterisk is a unique number for the HLA type. This number consists of up to four fields that are delimited by colons, each of which has its own meaning. The first field defines the allele group of the HLA type, which is determined by the serological antigens. The second field defines the proteins that are produced by the HLA type, which means that if two HLA types have the same first field but a different second field, the two HLA types produce different proteins. The third field is used to distinguish synonymous mutations of the HLA type, which means that if two HLA types have the same first and second fields but a different third field, the two HLA types produce the same protein but have different DNA sequences in the exons. The fourth filed is used to distinguish mutations in non-coding regions of the HLA types, which means that if two HLA types have the same first, second, and third fields but a different fourth field, the two HLA types have the same DNA sequences in the exons but different DNA sequences in the non-coding regions.

Some HLA types have an additional suffix such as N (Null), L (Low), S (Secreted), C (Cytoplasm), A (Aberrant), and Q (Questionable), which indicates the expression level of the HLA type compared with standard levels. The suffix N means that the HLA type is not expressed, which means that the proteins produced by the HLA type do not contribute the immune system. The suffix L

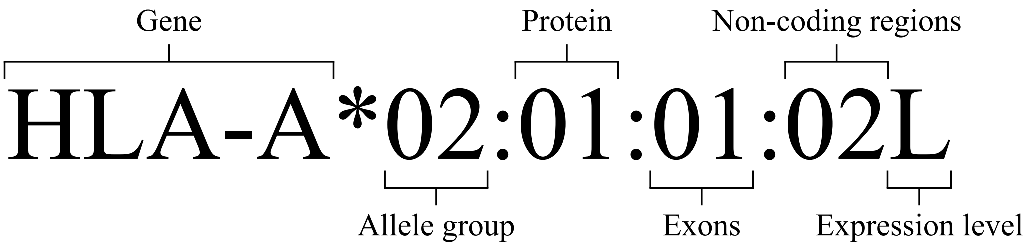


Figure 2.1: An example of an HLA type.

means that the expression level of the HLA type is low on the cell surface. The suffix S means that the HLA type produces soluble secreted proteins that are not expressed on the cell surface. The suffix C means that the proteins produced by the HLA type exist only in the cytoplasm but not on the cell surface. The suffix A means that there is some doubt whether the HLA type is expressed. The suffix Q means that the expression level of the HLA type is questionable because the HLA type has a mutation that changes the expression levels of other HLA types.

The prefix "HLA-" is sometimes omitted for simplicity.

2.2 NGS-based Sequence Data

Recently, NGS technologies have become an essential tool as they have generated a large amount of DNA and RNA sequence data with high throughput at low costs. In this section, we explain what is NGS data and fundamental tools to analyze NGS data. NGS data consists of sequenced DNA or RNA fragments from a sample. There are two methods to sequence DNA or RNA fragments: paired-ended sequencing and single-ended sequencing. Figure 2.2 shows the overview of paired-ended sequencing and single-ended sequencing. Paired-ended sequencing reads both ends of each DNA or RNA fragment (Figure 2.2(a)). On the other hand, single-ended sequencing reads either end of each DNA or RNA fragment (Figure 2.2(b)). These sequenced parts of DNA or RNA fragments are called sequence reads. Sequence reads are stored as character strings to a sequence data. Together with sequenced base pairs, the base qualities are also stored, which indicate how accurately the base pairs are sequenced. An instance of sequence data is shown in Figure 2.3. Each sequence read is written using four lines. The first line indicates the name of the sequence read. The second line indicates the sequenced base pairs. The third line is a delimiter. The fourth line is the base qualities. Note that the length of the base qualities is the same as that of the sequenced base pairs. The length of sequence reads is different by the model of the sequencer, which ranges tens of base pairs to tens of thousands of base pairs.

Sequence data is then processed according to the purpose of analysis. Gener-

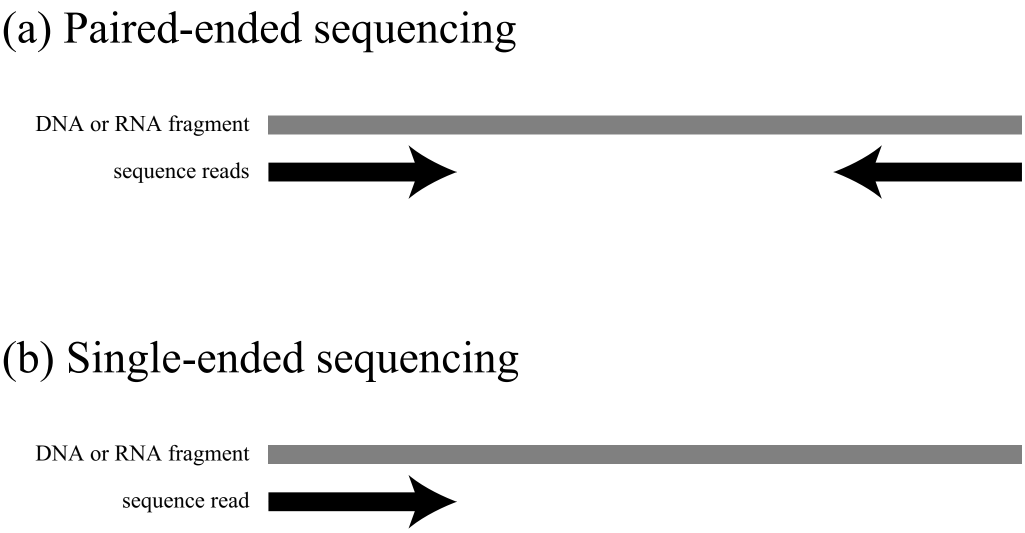


Figure 2.2: The overview of paired-ended sequencing and single-ended sequencing.

```
@Read1
ACCAGGTTACACCTTGATTTCTATAAAATC
+
GIG:BGECECHFFF<BECEC@CCDDCDCC@
@Read2
TGAACTACGCAATCTAATACTCG
+
HEFBDFFDCEABAACC:?BAC@;
...
```

Figure 2.3: An instance of sequence data.

ally, sequence reads are first aligned to a reference genome of the species. This process is called sequence alignment, which provides the information on where each read comes from in the genome. In recent years, a lot of sequence alignment methods have been developed [48, 61, 46, 49, 92, 87, 23, 43, 89, 37, 13, 50, 36]. Sequence alignment is a fundamental step of sequence data analysis.

One purpose of sequence data analysis is somatic mutation calling, that is, identifying somatic mutations of cancer patients. Somatic mutations can be called by comparing sequence data of normal and tumor cells, and detecting mutations that are seen only in tumor sequence data. Recently, several methods have been developed to achieve high accuracy in somatic mutation calling [47, 57, 44, 1, 39, 74, 71, 12, 9, 77]. Somatic mutation calling is essential to understanding the relationship between cancer and the immune system.

2.3 Markov Chain Monte Carlo Methods

In this section, we explain Markov chain Monte Carlo (MCMC) methods, which are algorithms that are used to sample from a probability distribution. In MCMC, a probability distribution of interest can be obtained as the limiting distribution of a Markov chain. There are several MCMC methods according to how the Markov chain is constructed. In the following sections, we introduce two MCMC methods, Gibbs sampling [18] and the Metropois-Hastings algorithm [58, 26]. We further describe parallel tempering [80, 19], which is also known as the replica exchange method, that is a technique for improving convergence of MCMC.

2.3.1 Gibbs Sampling

Gibbs sampling is an MCMC method where each parameter at the next step is sampled from its conditional distribution given all of the other parameters. The algorithm of Gibbs sampling is shown in Algorithm 2.1. Let $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_d)$ be parameters. First, all parameters are initialized. Although this initialization is theoretically arbitrary, optimization techniques are sometimes used to search better initial parameters for quick convergence of MCMC. Then, the i^{th} parameter θ_i^t at the time t is sampled from the full conditional $p(\theta_i \mid \theta_1^t, \ldots, \theta_{i-1}^t, \theta_{i+1}^{t-1}, \ldots, \theta_d^{t-1})$. This sampling is repeatedly conducted until the MCMC chain converges. Note that Gibbs sampling requires that each parameter can be easily sampled from the full conditional.

Algorithm 2.1 Gibbs sampling

```
Initialize $\theta_1^0, \cdots, \theta_d^0$
$t \leftarrow 1$
repeat
    for $i \leftarrow 1$ to $d$ do
        $\theta_i^t \sim p(\theta_i \mid \theta_1^t, \ldots, \theta_{i-1}^t, \theta_{i+1}^{t-1}, \ldots, \theta_d^{t-1})$
    end for
    $t \leftarrow t + 1$
until convergence
```

Obtained parameters from Gibbs sampling approximately follow the probability distribution $p(\boldsymbol{\theta})$ of interest. However, parameters sampled in the early period are not considered to follow the probability distribution because they are affected by the initial parameters. Therefore, parameters sampled in the early period are generally not used for the subsequent analysis. This period is called burn-in.

2.3.2 Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm is another MCMC method where parameters at the next step are sampled from a user-defined probability distribution, which is called a proposal distribution. The Metropolis-Hastings algorithm is shown in Algorithm 2.2. The Metropolis-Hastings algorithm can be used as long as the probability distribution of interest can be calculated, even if each parameter cannot be easily sampled from the full conditional. In the Metropolis-Hastings algorithm, new parameters are sampled from a proposal distribution $q^t(\boldsymbol{\theta^*} \mid \boldsymbol{\theta^{t-1}})$ at the time t. This algorithm is different from Gibbs sampling in that the new parameters are not always accepted. The acceptance ratio r is given

$$
\begin{aligned}
r &= \min(1, r^*) \\
r^* &= \frac{p(\boldsymbol{\theta^*})q^t(\boldsymbol{\theta^{t-1}} \mid \boldsymbol{\theta^*})}{p(\boldsymbol{\theta^{t-1}})q^t(\boldsymbol{\theta^*} \mid \boldsymbol{\theta^{t-1}})}.
\end{aligned}
$$

This acceptance process is required for MCMC convergence to the probability distribution of interest.

In the Metropolis-Hastings algorithm, proposal distributions determine the speed of MCMC convergence. Therefore, it is important to construct proposal

Algorithm 2.2 The Metropolis-Hastings algorithm

```
Initialize $\boldsymbol{\theta^0}$
$t \leftarrow 1$
repeat
    $\boldsymbol{\theta^*} \sim q^t(\boldsymbol{\theta^*} | \boldsymbol{\theta^{t-1}})$
    $r^* \leftarrow \dfrac{p(\boldsymbol{\theta^*})q^t(\boldsymbol{\theta^{t-1}} \mid \boldsymbol{\theta^*})}{p(\boldsymbol{\theta^{t-1}})q^t(\boldsymbol{\theta^*} \mid \boldsymbol{\theta^{t-1}})}$
    $r \leftarrow \min(1, r^*)$
    $\boldsymbol{\theta^t} \leftarrow \begin{cases} \boldsymbol{\theta^*} & \text{(with probability} r) \\ \boldsymbol{\theta^{t-1}} & \text{(otherwise)} \end{cases}$
    $t \leftarrow t + 1$
until convergence
```

distributions that lead to quick MCMC convergence. Also, as in the case of Gibbs sampling, parameters sampled in the burn-in period should be discarded.

2.3.3 Parallel Tempering

Parallel tempering is a scheme to be used for acceleration of MCMC convergence. Especially, it is effective when the probability distribution of interest is multimodal, where it is difficult for parameters to move from a local optimum to another. Parallel tempering runs multiple MCMC chains at different temperatures ($T_1 = 1, T_2, \ldots, T_m$). In other words, the i^{th} MCMC chain samples parameters from the probability distribution in proportion to $p(\boldsymbol{\theta})^{\frac{1}{T_i}}$. Each MCMC chain basically makes sampling independently of the other MCMC chains. Sometimes, parameters of two MCMC chains are exchanged using the Metropolis-Hastings algorithm. In the case of the i^{th} and j^{th} MCMC chains, the acceptance ratio r of the exchange of the parameters $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ is given by

$$\begin{aligned} r &= \min(1, r^*) \\ r^* &= \frac{p(\boldsymbol{\theta_j})^{\frac{1}{T_i}} p(\boldsymbol{\theta_i})^{\frac{1}{T_j}}}{p(\boldsymbol{\theta_i})^{\frac{1}{T_i}} p(\boldsymbol{\theta_j})^{\frac{1}{T_j}}} \\ &= \left(\frac{p(\boldsymbol{\theta_j})}{p(\boldsymbol{\theta_i})} \right)^{\frac{1}{T_i} - \frac{1}{T_j}} . \end{aligned}$$

After convergence, only the parameters sampled from the first MCMC chain are used for posterior inference.

At high temperature, the probability distribution becomes flat, and hence the influence of multimodality can be reduced. Through the exchange of parameters, parameters can move from a local optimum to another even at low temperature.

Chapter 3

Extraction, Classification, and Realignment of HLA Reads from Sequence Data

3.1 Overview

When we analyze HLA genes from WES, WGS, or RNA-seq data, we must first extract HLA reads from the sequence data since the sequence data contains reads that were produced by not only HLA genes but also other regions. In addition, for subsequent HLA analysis, we must judge which HLA gene produced each extracted HLA read. However, the extraction and classification of HLA reads must be carefully performed for various reasons. First, owing to the high polymorphism of HLA genes, it is insufficient to use only a human genome reference such as GRCh37 or GRCh38 because these human genome references have only one HLA type for each HLA locus. When a sample has an HLA type which is not the same as the human reference genome, sequence reads from HLA genes may not be correctly aligned to the reference genome, which can cause inaccurate extraction and classification of HLA reads. Therefore, a specific database of HLA types is required to deal with this problem. Second, HLA genes and HLA pseudogenes are paralogs, and are therefore quite similar. Figure 3.1 shows an example of the similarity among HLA genes and HLA pseudogenes. In such a situation, it is difficult to determine which HLA gene produced a read, although it is necessary for the following HLA genotyping. The first problem can be solved simply by using a database of HLA types. However, the second problem is not easy to solve, and a sophisticated method is required.

Most of existing methods cope with this problem by using the number of mismatches between each read and each HLA type. This approach works well when the quality of the sequence data is sufficiently high. In this case, we can judge that a read with a few of mismatches to an HLA type was not produced by the HLA type. However, if the sequence quality is low, we cannot determine

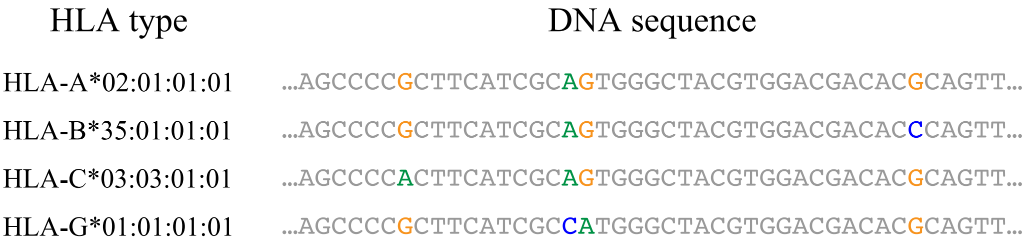

HLA type
DNA sequence
HLA-A*02:01:01:01 ...AGCCCCGCTTCATCGCAGTGGGCTACGTGGACGACACGCAGTT...
HLA-B*35:01:01:01 ...AGCCCCGCTTCATCGCAGTGGGCTACGTGGACGACACCCAGTT...
HLA-C*03:03:01:01 ...AGCCCCACTTCATCGCAGTGGGCTACGTGGACGACACGCAGTT...
HLA-G*01:01:01:01 ...AGCCCCGCTTCATCGCCATGGGCTACGTGGACGACACGCAGTT...

Figure 3.1: An example of the similarity among HLA genes and HLA pseudogenes.

whether the mismatches were caused because the read was not produced by the HLA type or because the read was really produced by the HLA type but has low base qualities. Therefore, we have developed a novel method to deal with this problem using alignment scores, which considers not only the number of mismatches but also base qualities at the mismatch positions.

The organization of this chapter is as follows. In section 3.2, we introduce some reference data used for our method. Then, in section 3.3, we describe our scoring method to extract and classify sequence reads from HLA genes.

The results in this chapter have been published as part of reference [28].

3.2 Reference Information on HLA Types

We use two databases as reference information on HLA types. The first one is a database that stores frequencies of HLA types, which is used to calculate prior probabilities of HLA types. The second one is a database that stores DNA sequences of HLA types, which is used as reference sequences in our methods.

3.2.1 Prior Probabilities of HLA Types

We first focus on prior information on HLA types which is used in our methods. We use the Allele Frequency Net Database [21] as prior information on frequencies of HLA types, which provides how many times each HLA type has been identified in studies done thus far. However, the database stores the frequencies of HLA types at various resolutions. Hence, we first infer the frequencies of HLA types at full resolution and then calculate prior probabilities of HLA types. Algorithms 3.1–3.6 show how this process is executed.

Algorithm 3.1 Calculate prior probabilities of HLA types

```
Input:
    $\boldsymbol{f}$: frequencies of HLA types
    $\gamma^{(\mathtt{f})}$: a pseudocount for frequencies of HLA types
Output:
    $\boldsymbol{p}$: prior probabilities of HLA types

function CalcPrior($\boldsymbol{f}, \gamma^{(\mathtt{f})}$)
    $F \leftarrow$ MakeFreqTrie($\boldsymbol{f}, \gamma^{(\mathtt{f})}$)
    $\boldsymbol{g} \leftarrow$ InferFreq($F$)
    $\boldsymbol{p} \leftarrow$ Freq2Prior($\boldsymbol{g}$)
    return $\boldsymbol{p}$
end function
```

Algorithm 3.2 Make a trie for frequencies of HLA types

Input:

$\boldsymbol{f}$: frequencies of HLA types

$\gamma^{(\mathtt{f})}$: a pseudocount for frequencies of HLA types

Output:

F: a frequency trie

```
function MakeFreqTrie($\boldsymbol{f}, \gamma^{(\mathtt{f})}$)
    $T^{(\mathtt{p})} \leftarrow$ a set of prefix HLA types
    $T^{(\mathtt{f})} \leftarrow$ a set of HLA types at full resolution
    $F \leftarrow$ a trie whose keys are HLA types in $T^{(\mathtt{p})}$ and values are 0
    for all $t \in T^{(\mathtt{p})}$ do
        Search$(F, t).freq \leftarrow f_t$
    end for
    for all $t \in T^{(\mathtt{f})}$ do
        Search$(F, t).freq \leftarrow$ Search$(F, t).freq + \gamma^{(\mathtt{f})}$
    end for
    return $F$
end function
```

Algorithm 3.3 Infer frequencies of HLA types at full resolution

Input:

F: a frequency trie

Output:

$\boldsymbol{g}$: inferred frequencies of HLA types at full resolution

```
function InferFreq($F$)
    CalcSubtrieFreq($F.root$)
    DistributeFreq($F.root$, 0)
    $T^{(\mathtt{f})} \leftarrow$ a set of HLA types at full resolution
    for all $t \in T^{(\mathtt{f})}$ do
        $g_t \leftarrow$ Search$(F, t).freq$
    end for
    return $\boldsymbol{g}$
end function
```

Algorithm 3.4 Calculate the total frequency of the HLA types in the subtrie

Input:

n: a node of a frequency trie

Output:

Nothing

```
function CalcSubtrieFreq($n$)
    $n.subtrie\_freq \leftarrow n.freq$
    for all $c \in n.children$ do
        CalcSubtrieFreq(c)
        $n.subtrie\_freq \leftarrow n.subtrie\_freq + c.subtrie\_freq$
    end for
end function
```

Algorithm 3.5 Distribute frequencies of HLA types to higher resolution

Input:
n: a node of a frequency trie
f: a distributed frequency from the parent node

Output:
Nothing

```
1: function DistributeFreq($n, f$)
2:     $n.inferred\_freq \leftarrow n.freq + f$
3:     $s \leftarrow n.subtrie\_freq - n.freq$
4:     for all $c \in n.children$ do
5:         $r \leftarrow \frac{c.subtrie\_freq}{s}$
6:         $d \leftarrow n.inferred\_freq \times r$
7:         DistributeFreq($c, d$)
8:     end for
9: end function
```

Algorithm 3.6 Convert frequencies into prior probabilities

Input:
$\boldsymbol{g}$: inferred frequencies of HLA types at full resolution

Output:
$\boldsymbol{p}$: prior probabilities of HLA types

```
function Freq2Prior($\boldsymbol{g}$)
    $\boldsymbol{p} \leftarrow \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|}$
    return $\boldsymbol{p}$
end function
```

First, we make a trie to infer the frequencies of HLA types at full resolution. This trie is made from the frequencies of HLA types that are obtained from the Allele Frequency Net Database using Algorithm 3.2. Figure 3.2 shows an instance of the constructed trie. In the trie, the keys are HLA types, and the depth corresponds to the resolution, which means that the leaves represent the HLA types at full resolution. Each node stores the frequency of the HLA type that corresponds to the node. We further use a hyperparameter of a pseudocount $\gamma^{(\mathtt{f})}$, which is added to frequencies of HLA types at full resolution. Next, the frequencies of HLA types at full resolution are inferred by Algorithm 3.3. We first calculate the total frequency of the HLA types in each subtrie by Algorithm 3.4. This total frequency of a node is equal to the sum of the frequency of the node and the total frequencies of the child subtries. The result of this algorithm in the instance is shown in Figure 3.3. Then, the frequency of each internal node is distributed to its children in proportion to the total frequencies of the child subtries by Algorithm 3.5. This distribution is made from the root to the leaves, that is, from lower resolution to higher resolution in order. Note that in Algorithm 3.5, s defined in line 3 is equal to the sum of the total frequencies of the child subtries. Figures 3.4–3.6 show the overview of the distribution in the instance. First, the frequency of the root should be distributed to its children, A*02 and A*24, but no distribution occurs from the root because the frequency of the root is always zero (Figure 3.4). Next, the frequency of the node A*02 is distributed to its children, A*02:01 and A*02:07 (Figure 3.5). Since the total frequencies of the subtries for A*02:01 and A*02:07 are 28 and 12, respectively, the frequency of the node A*02 is distributed to A*02:01 and A*02:07 at the ratio of 28:12 = 7:3, which means that the counts of 56 and 24 are added to A*02:01 and A*02:07. Then, the inferred frequency of A*02:01 becomes 70 (= 14 + 56), and the distribution of the inferred frequency is performed in the same way (Figure 3.6). By conducting the distribution recursively, we can obtain the inferred frequencies of all HLA types in the trie. Finally, prior probabilities of HLA types can be calculated from the inferred frequencies of HLA types at full resolution by Algorithm 3.6. The prior probabilities are assigned in proportion to the inferred frequencies.

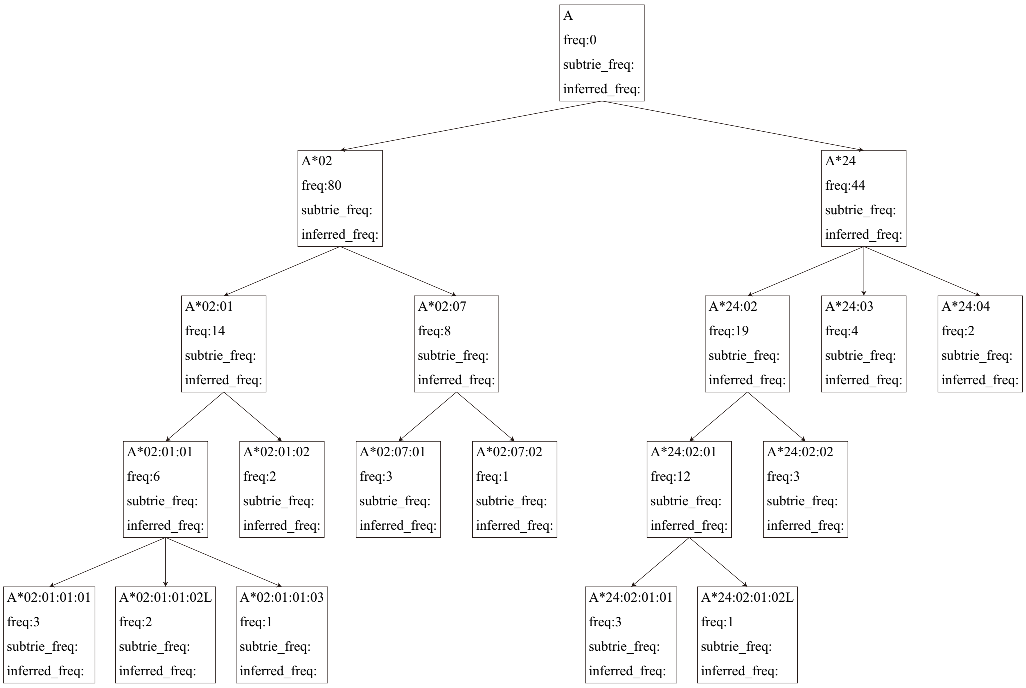


Figure 3.2: An instance of a frequency trie of the HLA-A gene.

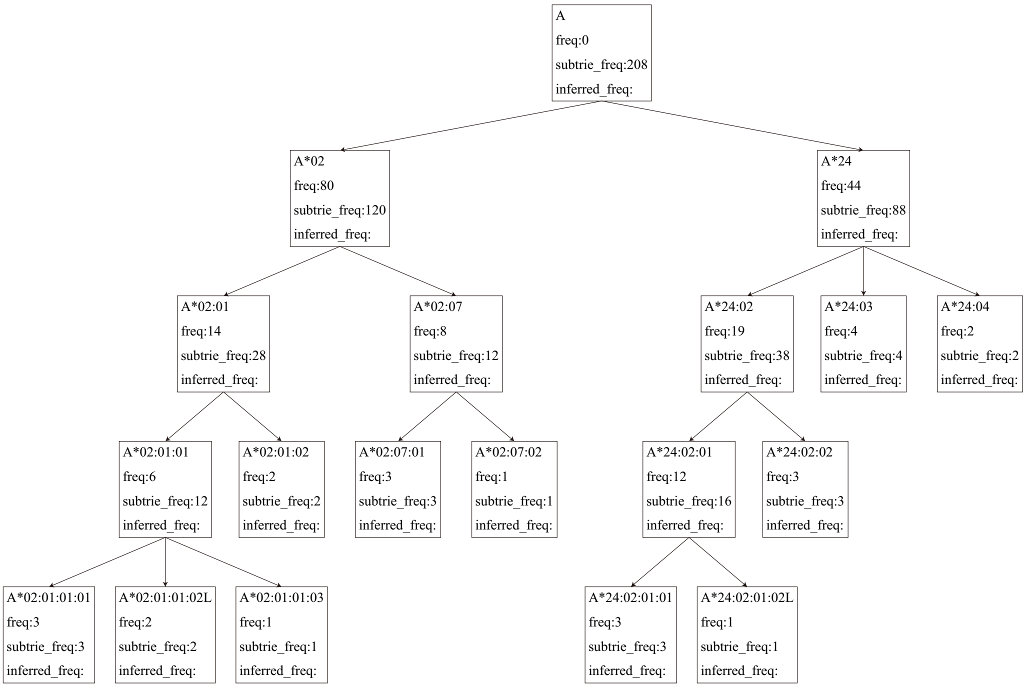


Figure 3.3: The frequency trie after calculation of the total frequencies for all subtries.

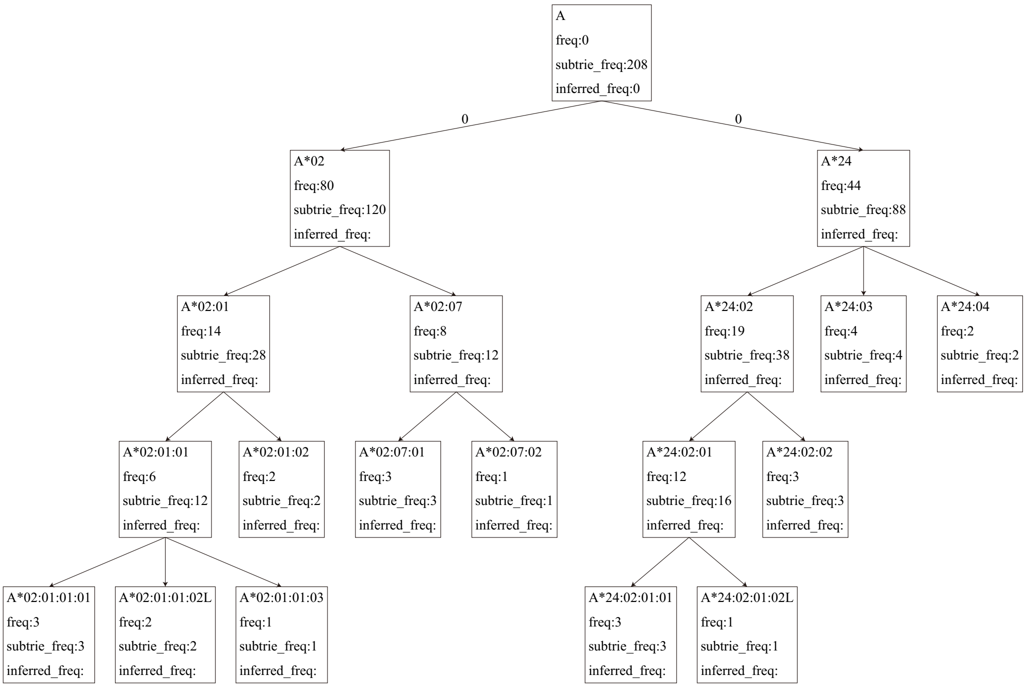


Figure 3.4: Distribution of the frequency of the root.

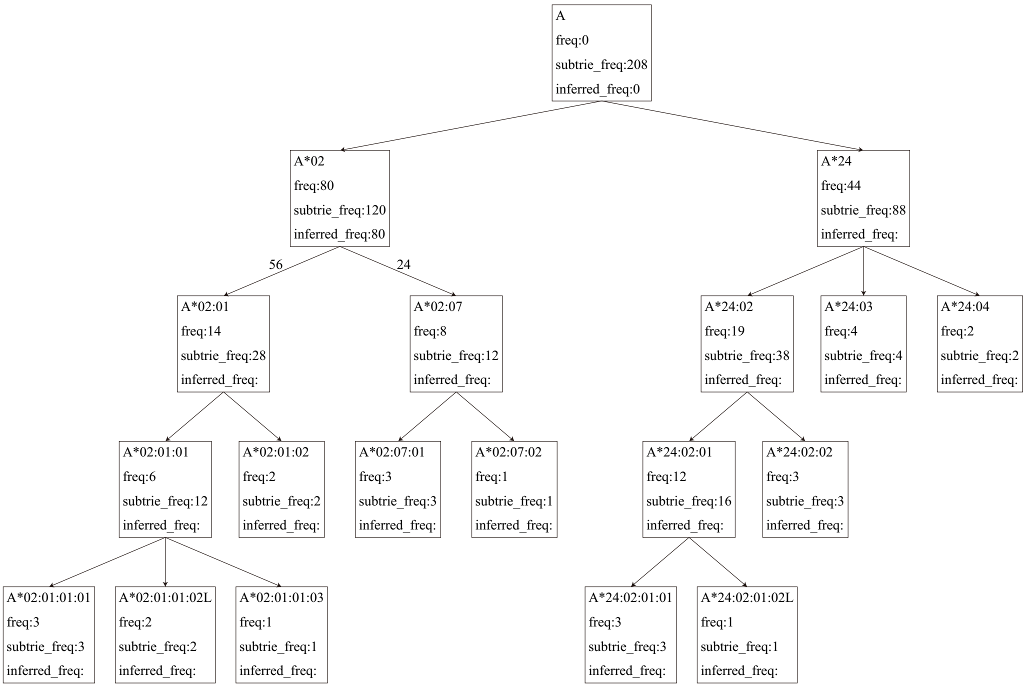


Figure 3.5: Distribution of the frequency of the node A*02.

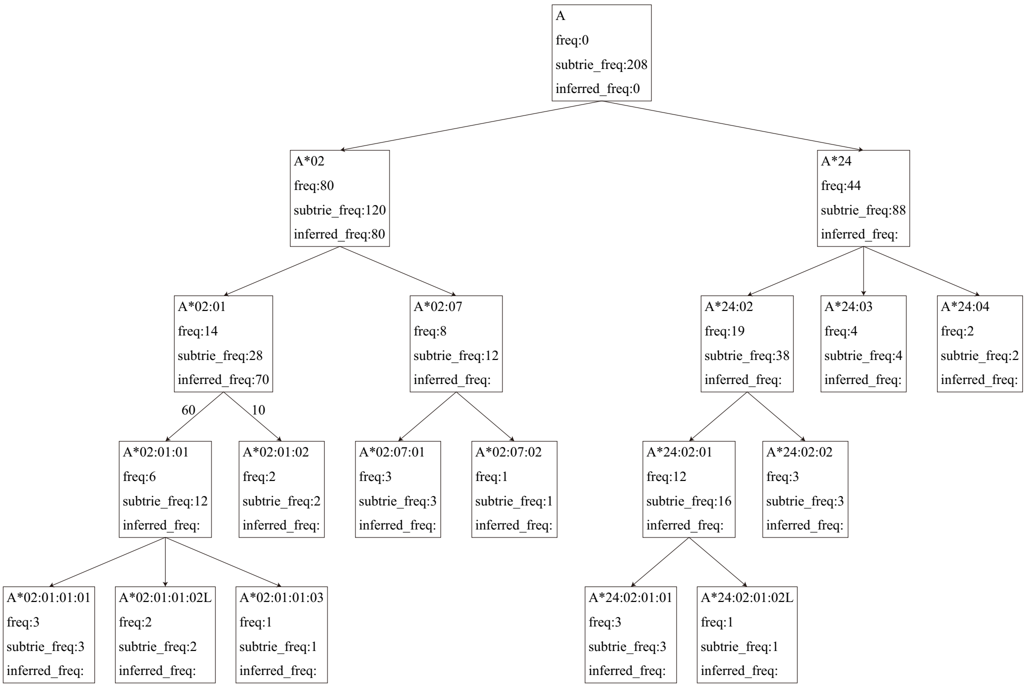


Figure 3.6: Distribution of the frequency of the node A*02:01.

3.2.2 Reference Sequences of HLA Types

Next, we focus on reference sequences of HLA types in our methods. We use the IPD-IMGT/HLA Database [69], from which we can obtain the DNA sequences of HLA types. However, most of the HLA DNA sequences that are registered in the database are incomplete, which have several exons and introns whose

DNA sequences are unavailable and unknown. Therefore, we first impute the unknown bases so that we can judge whether a read is derived from an HLA gene even if the sample has an incomplete HLA type. To this end, we use multiple sequence alignments (MSAs) of HLA types, which can be also obtained from the IPD-IMGT/HLA Database. The IPD-IMGT/HLA Database provides two types of MSAs for each HLA locus: an MSA at the genomic level and an MSA at the exonic level. The genomic MSA is a sequence alignment of both exons and introns for each HLA locus but includes only a part of the HLA types. On the other hand, the exonic MSA is a sequence alignment of only exons but includes all of the HLA types. We first integrate the genomic MSA and the exonic MSA into an MSA that is consistent with the two MSAs. The integrated MSA has the same exonic part as the exonic MSA. Also, it has the same intronic part as the genomic MSA for the HLA types that are included in the genomic MSA. For the HLA types that are not included in the genomic MSA, the intron sequences are unknown, and hence the intronic part is filled with `N`, which indicates an unknown base. Then, we impute the unknown bases by replacing each `N` with the base of the most similar HLA type. The similarity is measured by the Hamming distance in the combined MSA. If there are multiple HLA types that have the smallest Hamming distance, the HLA type with the highest prior probability is used. Thus, HLA reference sequences with no `N`s can be obtained.

3.3 Extraction, Classification, and Realignment of HLA Reads

3.3.1 Rough Extraction of HLA Reads

Since sequence data contains reads from regions other than HLA genes, we first conduct a rough extraction of HLA reads. This extraction consists of the following two steps: aligning all reads in the sequence data to a human reference genome and extracting reads that are aligned to the HLA region. In the alignment, we use BWA-MEM [46] as the aligner and GRCh37 as the human reference genome. BWA-MEM is a popular alignment tool for effectively aligning reads to a long reference with Burrows–Wheeler Transform [6], which requires linear time with respect to the length of the read and independent of the length of the reference genome. Then, sequence reads are filtered by extracting the HLA region, which is defined by the interval from the 28,477,797th base to the 33,448,354th base on chromosome 6 for GRCh37. This region covers all of the HLA class I genes and HLA class II genes, and most of HLA reads are considered to be included in the extracted reads.

3.3.2 Scoring Method for Extraction and Classification of HLA Reads

Next, the extracted reads are mapped to reference sequences of HLA types of all HLA loci, which are constructed in Section 3.2.2, using BWA-MEM. Since it is possible that a read is aligned to multiple HLA types, we use BWA-MEM with the -a option, which provides information on all identified alignments. Also, by giving a sufficiently large clipping penalty, we do not allow that reads are not aligned to the reference sequences from end to end. Then, each aligned read is classified based on whether or not the HLA genes produced the read, and if so, which specific gene was involved. Figure 3.7 shows this procedure. For each aligned read and each HLA type, the HLA read score (HR score) is calculated, which quantifies the likelihood that the read comes from the HLA type. Based on the calculated HR scores, it is determined whether or not the read comes from

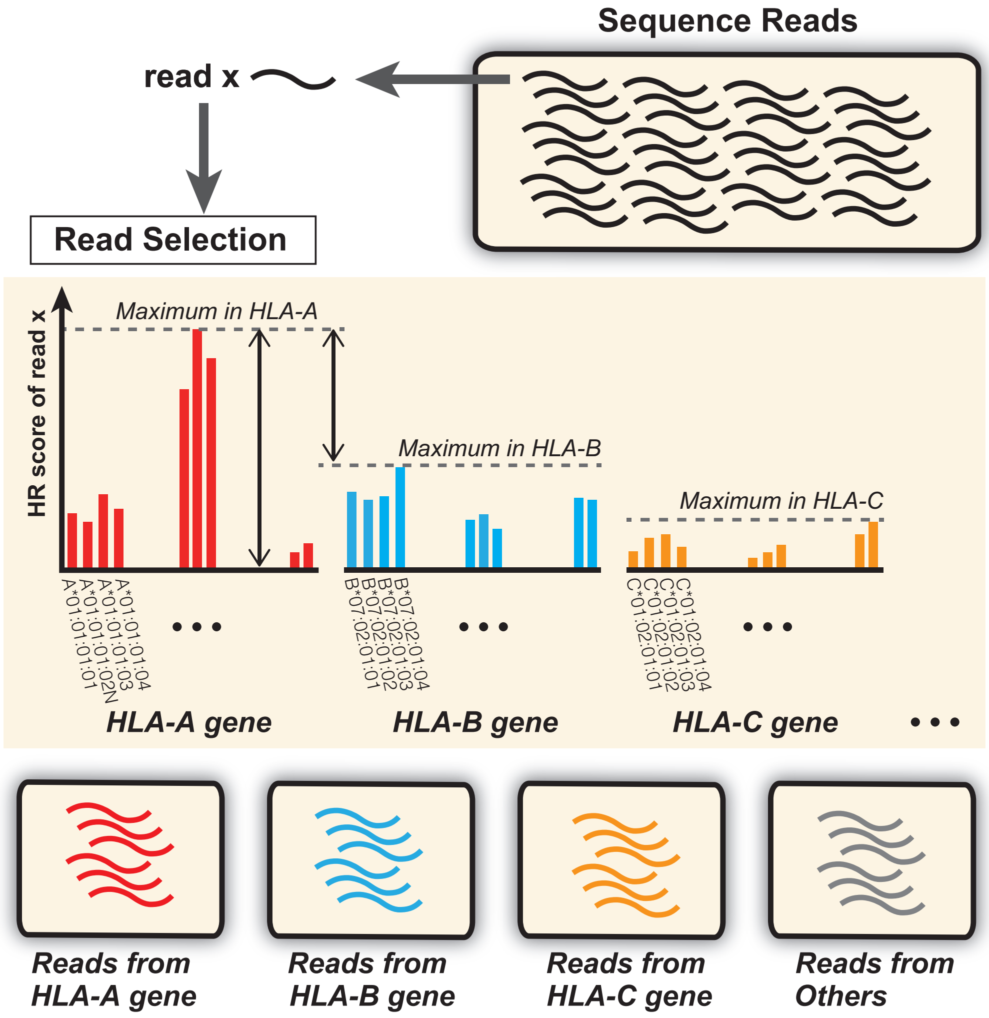


Figure 3.7: The overview of our scoring approach. For each read and each HLA type, the HLA read score (HR score) is calculated, which quantifies the likelihood that the read comes from the HLA type. Based on the calculated HR scores, it is determined whether or not the read comes from a specific HLA gene.

a certain HLA gene. Specifically, HR scores are calculated as follows. Let x_i be the i^{th} read pair that consists of two single reads, $x_{i,1}$ and $x_{i,2}$. If the i^{th} read is single-ended, x_i consists of one read, $x_{i,1}$. In addition, t_k is defined as the k^{th} HLA type. Then, for each read $x_{i,j}$ and each HLA type t_k, we calculate the HR score $s_{i,j,k}$, which indicates how likely the read $x_{i,j}$ was produced by the HLA type t_k , based on the alignment information. If the read $x_{i,j}$ is unmapped to the HLA type t_k, then the HR score $s_{i,j,k}$ 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. Besides, we define $\tilde{b}_{i,j,k,n}$ as the Phred base quality of $\tilde{x}_{i,j,k,n}$. Then, 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

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

Using the above definitions, the HR score $s_{i,j,k}$ is given by

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

where

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

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

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 the reward for one base. By contrast, $s^{(\mathtt{p})}_{i,j,k,n}$ is a penalty for a mismatch between the read and the HLA type, and $\alpha^{(\mathtt{d,o})}$, $\alpha^{(\mathtt{d,e})}$, $\alpha^{(\mathtt{i,o})}$, $\alpha^{(\mathtt{i,e})}$, 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. These formulas suggest that a longer read with fewer mismatches, insertions, and deletions achieves a higher HR score. Also, a mismatch with a lower base quality is preferable and gives a smaller penalty.

Then, by using the calculated HR scores, we judge whether the read pair was produced by the HLA genes. For each read $x_{i,j}$ and each HLA locus l, the score $s^*_{i,j,l}$ is defined by

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

where T_l is a set of HLA types of the HLA locus l. Then, we define the score $s^*_{i,l}$ for the read pair x_i and the HLA locus l by

$$s^*_{i,l} = \sum_j s^*_{i,j,l}.$$

The score $s^*_{i,l}$ indicates the possibility that the read pair x_i was produced by the HLA locus l.

Based on the calculated scores, we judge whether each read should be used for HLA genotyping for a specific HLA locus. When x_i is a paired-ended read, the read pair is used for genotyping the HLA locus l if the following two criteria are satisfied:

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

Here, $\theta^{(\mathtt{p,s})}$ is a hyperparameter of a threshold for the maximum HR score of the locus, and $\theta^{(\mathtt{p,d})}$ is a hyperparameter of a threshold for the difference between the maximum HR scores of the locus and other loci. The first condition guarantees that the read pair x_i is well aligned to the HLA locus l. This condition is necessary to collect reads that were likely to be produced by the locus. The second condition guarantees that the read pair x_i is not well aligned to HLA loci other than l. This condition is necessary because of the high similarity among HLA genes and HLA pseudogenes. Even if the first condition is satisfied, we cannot decide whether the read pair x_i was produced by the HLA locus l or another HLA locus l' if the two scores $s^*_{i,l}$ and $s^*_{i,l'}$ have similar values. Therefore, the second condition is needed to exclude reads that might be produced by other HLA loci. On the other hand, 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

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

Note that $\theta^{(\mathtt{s},\mathtt{s})}$ and $\theta^{(\mathtt{s},\mathtt{d})}$ should be larger than or equal to $\theta^{(\mathtt{p},\mathtt{s})}$ and $\theta^{(\mathtt{p},\mathtt{d})}$, respectively. This is because even if a paired-ended read has the same length as a single-ended read, the paired-ended read can cover a broader region than the single-ended read, which means that paired-ended reads are more unlikely to be well aligned to HLA loci other than the HLA locus that produced them. Thus, HLA reads are extracted and classified into HLA loci.

3.3.3 Realignment of HLA Reads

In this section, we focus on realignment of the extracted HLA reads that are aligned to the HLA locus l. This realignment is necessary for the subsequent HLA genotyping step mentioned in the following chapters. For each read $x_{i,j}$, the realignment is performed using the HLA type that achieves the best HR score, whose index is given by

$$k^* = \operatorname*{arg\,max}_{k : t_k \in T_l} s_{i,j,k}.$$

If there are multiple HLA types that achieve the best HR score, the HLA type with the highest prior probability is adopted. Then, the read $x_{i,j}$ is realigned to the HLA type t_{k^*}. Specifically, the realigned read $\hat{x}_{i,j}$ is obtained by aligning $x_{i,j}$ to the aligned sequence $\hat{t}_{k^*}$ in the combined MSA, which was constructed in Section 3.2.2, based on 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 $\hat{t}_{k^*}$.

Chapter 4

Bayesian Approach for HLA Genotyping from Whole Genome Sequence Data

4.1 Overview

HLA genotyping is essential to a lot of research fields related to immunology. Recent advances in sequencing technology have facilitated HLA genotyping from NGS data such as WES, WGS, and RNA-seq data. A lot of methods have been developed and have achieved high accuracy for WES and RNA-seq data. However, HLA genotyping from WGS data is still challenging because the data is relatively shallow.

In this chapter, we introduce a new Bayesian method, called ALPHLARD, that accurately determines HLA genotypes for each of the HLA-A, HLA-B, HLA-C, HLA-DPA1, HLA-DPB1, HLA-DQA1 HLA-DQB1, and HLA-DRB1 genes at third field resolution from WGS data as well as WES data. ALPHLARD conducts HLA genotyping for each HLA locus independently by using reads that were classified into the HLA locus in Chapter 3. We can estimate the HLA genotype of a sample by calculating the posterior distribution of the Bayesian model. In ALPHLARD, the posterior distribution is calculated using MCMC. ALPHLARD can also detect germline variants, which means that ALPHLARD can identify novel HLA types that are not stored in the IPD-IMGT/HLA Database. In addition to germline variants, ALPHLARD can call somatic mutations by using paired normal and tumor sequence data.

The organization of this chapter is as follows. First, in Section 4.2, we explain three existing methods, OptiType [81], PHLAT [3], and HLA-VBSeq [62], that are used for performance comparison. Then, we introduce a specific Bayesian model of ALPHLARD in Section 4.3. In Section 4.4, we explain how parameters are sampled from the Bayesian model. Lastly, we show some experimental results in 4.5.

The results in this chapter have been published as reference [28].

4.2 Related Work

4.2.1 OptiType

OptiType [81] is a method for HLA genotyping at second field resolution from WES, WGS, and RNA-seq data. OptiType identifies HLA genotypes simultaneously for the HLA-A, HLA-B, HLA-C, HLA-G, HLA=H, and HLA-J genes on the assumption that the HLA genotypes to which the largest number of reads are mapped is correct. In OptiType, this problem is formulated as integer linear programming, which is known as an NP-hard problem. Bauer *et al.* has reported that OptiType achieves the best accuracy for HLA class I genotyping from all of WES (98%), WGS (71%), and RNA-seq (99%) data [4]. However, there are some HLA types that cannot be identified by OptiType because OptiType uses only exons 2 and 3 in the HLA reference sequences. In addition, OptiType outputs the results of HLA class I genes only.

4.2.2 PHLAT

PHLAT [3] performs HLA genotyping at full resolution from WES and RNA-seq data. PHLAT predicts HLA genotypes for the HLA-A, HLA-B, HLA-C, HLA-DQA1, HLA-DQB1, and HLA-DRB1 genes. HLA genotyping in PHLAT consists of two steps. At the first step, PHLAT narrows down candidate HLA types by counting the number of mapped reads based on the results of alignment. At the second step, PHLAT searches for the best pair of the candidate HLA types based on the likelihoods. Bauer *et al.* has reported that PHLAT achieves the best accuracy for HLA genotyping from WES (73%) and RNA-seq (81%) data [4].

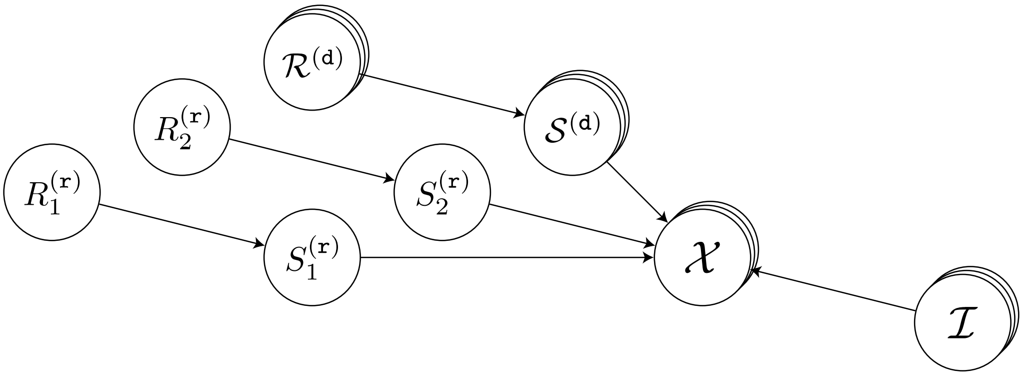
4.2.3 HLA-VBSeq

HLA-VBSeq [62] is a method for HLA genotyping at full resolution from WES, WGS, and RNA-seq data. HLA-VBSeq determines HLA genotypes for the HLA-A, HLA-B, HLA-C, HLA-DQA1, HLA-DQB1 and HLA-DRB1 genes by default, but it can be easily extended to other HLA genes such as HLA-DPA1 and HLA-DPB1. HLA-VBSeq uses a Bayesian model to estimate the HLA genotype and calculates the posterior distribution using variational Bayesian inference [34], which produces an approximation of the posterior distribution by factorizing it. Bauer *et al.* has reported that HLA-VBSeq achieves the best accuracy for HLA genotyping from WGS (52%) [4].

4.3 Bayesian Model for HLA Genotyping

Hereafter in this chapter, we fix an HLA locus of interest to be genotyped. We first introduce the statistical representation of our Bayesian model ALPHLARD. Figure 4.1 shows the graphical model of ALPHLARD. Let x_i be the i^{th} realigned paired reads or unpaired read obtained in Chapter 3. Note that we use x_i instead of $\hat{x}_i$ for simplicity. We define $x_{i,j,n}$ as the n^{th} base or gap of the read $x_{i,j}$, and $p_{i,j,n}$ as the mismatch probability of $x_{i,j,n}$. Note that the first position of each realigned read is not the beginning of the read but rather the beginning of the combined MSA, and $x_{i,j,n}$ and $p_{i,j,n}$ are undefined if the n^{th} position is not covered by the read.

Suppose that $R_1^{(\mathtt{r})}$ and $R_2^{(\mathtt{r})}$ are HLA types of the sample. We also define $S_1^{(\mathtt{r})}$ and $S_2^{(\mathtt{r})}$ as HLA sequences of the sample, which are introduced because the sample might have a novel HLA type which is not registered in the IPD-IMGT/HLA database. In addition to the parameters that are defined above, we introduce decoy parameters for robust inference of HLA genotyping. Let $R_1^{(\mathtt{d})}, \ldots, R_{\nu^{(\mathtt{d})}}^{(\mathtt{d})}$ be decoy HLA types, where $\nu^{(\mathtt{d})}$ is a hyperparameter of the number of the decoy parameters. Similarly, we define $S_1^{(\mathtt{d})}, \ldots, S_{\nu^{(\mathtt{d})}}^{(\mathtt{d})}$ as decoy HLA sequences. These parameters are essential to make a robust inference because their presence can reduce the influence of misclassified reads in Chapter 3 that were actually produced by other HLA genes and HLA pseudogenes. For convenience, we sometimes use $(R_1, R_2, R_3, \ldots, R_{\nu^{(\mathtt{d})}+2})$ and $(S_1, S_2, S_3, \ldots, S_{\nu^{(\mathtt{d})}+2})$



$R_1^{(\mathtt{r})}$, $R_2^{(\mathtt{r})}$: HLA types

$\mathcal{R}^{(\mathtt{d})}$: decoy HLA types

$S_1^{(\mathtt{r})}$, $S_2^{(\mathtt{r})}$: HLA sequences

$\mathcal{S}^{(\mathtt{d})}$: decoy HLA sequences

$\mathcal{X}$: realigned read pairs

$\mathcal{I}$: variables that indicate which HLA sequence produced each read pair

Figure 4.1: Graphical representation of ALPHLARD.

instead of $(R_1^{(\mathtt{r})}, R_2^{(\mathtt{r})}, R_1^{(\mathtt{d})}, \dots, R_{\nu^{(\mathtt{d})}}^{(\mathtt{d})})$ and $(S_1^{(\mathtt{r})}, S_2^{(\mathtt{r})}, S_1^{(\mathtt{d})}, \dots, S_{\nu^{(\mathtt{d})}}^{(\mathtt{d})})$, respectively. We also denote the n^{th} base or gap of R_m in the combined MSA by $R_{m,n}$, and the n^{th} base or gap of S_m by $S_{m,n}$. Note that all of the R_m's and S_m's have the same sequence length as aligned sequences in the combined MSA, which we denote by N. We next define I_i as a parameter to indicate which HLA sequence produced the read pair x_i; that is, $I_i = m$ means that x_i was produced by S_m. Therefore, the influence of misclassified reads from other HLA genes and HLA pseudogenes can be ignored by assigning the corresponding indicator variables to decoy HLA sequences.

Then, the posterior probability of the parameters is given by

$$p(\mathcal{R}, \mathcal{S}, \mathcal{I} \mid \mathcal{X}) \propto p(\mathcal{X} \mid \mathcal{S}, \mathcal{I}) p(\mathcal{S} \mid \mathcal{R}) p(\mathcal{R}) p(\mathcal{I}),$$

where $\mathcal{R} = (R_1, \dots, R_{\nu^{(\mathtt{d})}+2})$, $\mathcal{S} = (S_1, \dots, S_{\nu^{(\mathtt{d})}+2})$, $\mathcal{I} = (I_1, I_2, \dots)$, and $\mathcal{X} = (x_1, x_2, \dots)$.

$p(\mathcal{X} \mid \mathcal{S}, \mathcal{I})$ is the likelihood of realigned read pairs and is defined by

$$\begin{aligned} p(\mathcal{X} \mid \mathcal{S}, \mathcal{I}) &= \prod_i p(x_i \mid S_{I_i}) \\ &= \prod_i \prod_j p(x_{i,j} \mid S_{I_i}) \\ &= \prod_i \prod_j \prod_n p(x_{i,j,n} \mid S_{I_i,n}), \end{aligned}$$

where

$$\begin{aligned}
&p(x_{i,j,n}|S_{m,n} \in B) \\
&\quad = \begin{cases}
(1-\pi^{(\mathtt{e,d})})(1-\pi^{(\mathtt{e,N})})(1-p_{i,j,n}) & (\text{if } x_{i,j,n} = S_{m,n}) \\
(1-\pi^{(\mathtt{e,d})})(1-\pi^{(\mathtt{e,N})})\frac{p_{i,j,n}}{3} & (\text{if } x_{i,j,n} \in B \text{ and } x_{i,j,n} \neq S_{m,n}) \\
(1-\pi^{(\mathtt{e,d})})\pi^{(\mathtt{e,N})} & (\text{if } x_{i,j,n} = \mathtt{N}) \\
\pi^{(\mathtt{e,d})} & (\text{if } x_{i,j,n} = \mathtt{-})
\end{cases}, \\
&p(x_{i,j,n} \mid S_{m,n} = \mathtt{-}) \\
&\quad = \begin{cases}
\pi^{(\mathtt{e,i})}(1-\pi^{(\mathtt{e,N})})\frac{1}{4} & (\text{if } x_{i,j,n} \in B) \\
\pi^{(\mathtt{e,i})}\pi^{(\mathtt{e,N})} & (\text{if } x_{i,j,n} = \mathtt{N}) \\
1-\pi^{(\mathtt{e,i})} & (\text{if } x_{i,j,n} = \mathtt{-})
\end{cases}, \\
&p(x_{i,j,n} \mid S_{m,n} = \mathtt{N}) \\
&\quad = \begin{cases}
(1-\pi^{(\mathtt{e,N})})\frac{1}{5} & (\text{if } x_{i,j,n} \in B \text{ or } x_{i,j,n}\mathtt{-}) \\
\pi^{(\mathtt{e,N})} & (\text{if } x_{i,j,n} = \mathtt{N})
\end{cases}.
\end{aligned}$$

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

$p(\mathcal{S} \mid \mathcal{R})$ is the prior probability of HLA sequences and is defined by

$$\begin{aligned}
p(\mathcal{S} \mid \mathcal{R}) &= \left(\prod_m p(S_m^{(\mathtt{r})} \mid R_m^{(\mathtt{r})})\right)\left(\prod_m p(S_m^{(\mathtt{d})} \mid R_m^{(\mathtt{d})})\right) \\
&= \left(\prod_m \prod_n p(S_{m,n}^{(\mathtt{r})} \mid R_{m,n}^{(\mathtt{r})})\right)\left(\prod_m \prod_n p(S_{m,n}^{(\mathtt{r})} \mid R_{m,n}^{(\mathtt{r})})\right),
\end{aligned}$$

where

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

$$p(S_{m,n}^{(\mathtt{r})} \mid R_{m,n}^{(\mathtt{r})} \in B, R_{m,n}^{(\mathtt{r})} \text{ is imputed})$$
$$= \begin{cases} (1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{d})})(1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{s})}) & (\text{if } S_{m,n}^{(\mathtt{r})} = R_{m,n}^{(\mathtt{r})}) \\ (1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{d})})\frac{\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{s})}}{3} & \left(\begin{array}{l} \text{if } S_{m,n}^{(\mathtt{r})} \in B \\ \quad \text{and } S_{m,n}^{(\mathtt{r})} \neq R_{m,n}^{(\mathtt{r})} \end{array} \right) \\ (1-\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{N})})\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{d})} & (\text{if } S_{m,n}^{(\mathtt{r})} = \texttt{-}) \\ \pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{N})} & (\text{if } S_{m,n}^{(\mathtt{r})} = \mathtt{N}) \end{cases},$$

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

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

$$p(S_{m,n}^{(\mathtt{d})} \mid R_{m,n}^{(\mathtt{d})} \in B, R_{m,n}^{(\mathtt{d})} \text{ is original})$$
$$= \begin{cases} (1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{d})})(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{s})}) & (\text{if } S_{m,n}^{(\mathtt{d})} = R_{m,n}^{(\mathtt{d})}) \\ (1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{d})})\frac{\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{s})}}{3} & \left(\begin{array}{l} \text{if } S_{m,n}^{(\mathtt{d})} \in B \\ \quad \text{and } S_{m,n}^{(\mathtt{d})} \neq R_{m,n}^{(\mathtt{d})} \end{array} \right) \\ (1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{N})})\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{d})} & (\text{if } S_{m,n}^{(\mathtt{d})} = \texttt{-}) \\ \pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{N})} & (\text{if } S_{m,n}^{(\mathtt{d})} = \mathtt{N}) \end{cases},$$

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

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

$$
\begin{aligned}
&p(S^{(\mathtt{d})}_{m,n} \mid R^{(\mathtt{d})}_{m,n} \in B, R^{(\mathtt{d})}_{m,n} \text{ is imputed}) \\
&= \begin{cases}
(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{d})})(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{s})}) & (\text{if } S^{(\mathtt{d})}_{m,n} = R^{(\mathtt{d})}_{m,n}) \\
(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{d})})\frac{\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{s})}}{3} & \left(\begin{array}{c}\text{if } S^{(\mathtt{d})}_{m,n} \in B \\ \quad\text{and } S^{(\mathtt{d})}_{m,n} \neq R^{(\mathtt{d})}_{m,n}\end{array}\right) \\
(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})})\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{d})} & (\text{if } S^{(\mathtt{d})}_{m,n} = \texttt{-}) \\
\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})} & (\text{if } S^{(\mathtt{d})}_{m,n} = \mathtt{N})
\end{cases}, \\
&p(S^{(\mathtt{d})}_{m,n} \mid R^{(\mathtt{d})}_{m,n} = \texttt{-}, R^{(\mathtt{d})}_{m,n} \text{ is imputed}) \\
&= \begin{cases}
(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})})\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{i})}\frac{1}{4} & (\text{if } S^{(\mathtt{d})}_{m,n} \in B) \\
(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})})(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{i})}) & (\text{if } S^{(\mathtt{d})}_{m,n} = \texttt{-}) \\
\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})} & (\text{if } S^{(\mathtt{d})}_{m,n} = \mathtt{N})
\end{cases}, \\
&p(S^{(\mathtt{d})}_{m,n} \mid R^{(\mathtt{d})}_{m,n} = R^{(\mathtt{d})}_{m,n} \text{ is imputed}) \\
&= \begin{cases}
(1-\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})})\frac{1}{5} & (\text{if } S^{(\mathtt{d})}_{m,n} \in B \text{ or } S^{(\mathtt{d})}_{m,n} = \texttt{-}) \\
\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})} & (\text{if } S^{(\mathtt{d})}_{m,n} = \mathtt{N})
\end{cases}.
\end{aligned}
$$

Here, $\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{s})}$, $\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{d})}$, $\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{i})}$, $\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{N})}$ are hyperparameters of the probabilities of a germline substitution, a germline deletion, a germline insertion, and an N, respectively, in a non-decoy HLA sequence at the position where the reference is an original base. Similarly, $\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{s})}$, $\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{d})}$, $\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{i})}$, $\pi^{(\mathtt{g},\mathtt{r},\mathtt{i},\mathtt{N})}$ are hyperparameters of the probabilities of a germline substitution, a germline deletion, a germline insertion, and an N, respectively, in a non-decoy HLA sequence at the position where the reference is an imputed base. Also, $\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{s})}$, $\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{d})}$, $\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{i})}$, $\pi^{(\mathtt{g},\mathtt{d},\mathtt{o},\mathtt{N})}$ are hyperparameters of the probabilities of a germline substitution, a germline deletion, a germline insertion, and an N, respectively, in a decoy HLA sequence at the position where the reference is an original base. Lastly, $\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{s})}$, $\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{d})}$, $\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{i})}$, $\pi^{(\mathtt{g},\mathtt{d},\mathtt{i},\mathtt{N})}$ are hyperparameters of the probabilities of a germline substitution, a germline deletion, a germline insertion, and an N, respectively, in a decoy HLA sequence at the position where the reference is an imputed base. These hyperparameters determine how likely germline mutations in HLA genes are to occur. The probabilities for an imputed reference base should be larger than or equal to those for an original base to reduce the influence of misimputation. In addition, the probabilities for a decoy HLA sequence should also be larger than or equal to those for a non-decoy HLA sequence to achieve robustness against misclassified reads. Moreover, $S_{m,n}$ can be N, and tends to be N when $S_{m,n}$ cannot be uniquely determined, which can occur when the realigned read pairs $\mathcal{X}$ includes a number of misclassified reads from other HLA genes and HLA pseudogenes that cannot be ignored. In this case, allowing $S_{m,n}$ to be N can make HLA genotyping robuster.

$p(\mathcal{R})$ is the prior probability of HLA types and is defined by

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

Here, $p(R^{(\mathtt{r})}_m)$ is the prior probability of the HLA type and is given from the result of Section 3.2.1. On the other hand, $p(R^{(\mathtt{d})}_m)$ is the prior probability of the decoy HLA type, which we assume as constant.

$p(\mathcal{I})$ is the prior probability of indicator variables and is defined by

$$
p(\mathcal{I}) = \prod_i p(I_i),
$$

where

$$p(I_i) = \begin{cases} \frac{1-\pi^{(\mathtt{d})}}{2} & (\text{if } I_i \in \{1,2\}) \\ \frac{\pi^{(\mathtt{d})}}{\nu^{(\mathtt{d})}} & (\text{if } I_i \in \{3,\ldots,\nu^d+2\}) \end{cases}.$$

Here, $\pi^{(\mathtt{d})}$ is a hyperparameter that reflects how many misclassified reads are included in the realigned read pairs $\mathcal{X}$. Note that $I_i \in \{1,2\}$ means that the read pair x_i is assigned to a non-decoy HLA sequence, and $I_i \in \{3,\ldots,\nu^d+2\}$ means that x_i is assigned to a decoy HLA sequence.

4.4 MCMC Sampling of Parameters

To calculate the posterior distribution that is mentioned above, we use MCMC methods to sample parameters from the posterior distribution with parallel tempering to make the parameter sampling efficient. Gibbs sampling is mainly used to sample each parameter for local search. In addition, we periodically use the Metropolis-Hastings algorithm that allows parameters to move from mode to mode. In the following sections, we introduce how parameter sampling is conducted.

4.4.1 Gibbs Sampling of HLA Types $\mathcal{R}$

The conditional distribution of HLA types is given by

$$\begin{aligned} p(\mathcal{R} \mid \mathcal{S}, \mathcal{I}, \mathcal{X}) &\propto p(\mathcal{S} \mid \mathcal{R}) p(\mathcal{R}) \\ &= \prod_m \left(\prod_n p(S_{m,n} \mid R_{m,n}) \right) p(R_m). \end{aligned}$$

Therefore, each HLA type R_m can be sampled independently from the probability distribution in proportion to

$$p(S_m \mid R_m) p(R_m),$$

using Gibbs sampling. The Gibbs sampling of each HLA type is shown in Algorithm 4.1.

4.4.2 Gibbs Sampling of HLA Sequences $\mathcal{S}$

The conditional distribution of HLA sequences is given by

$$\begin{aligned} p(\mathcal{S} \mid \mathcal{R}, \mathcal{I}, \mathcal{X}) &\propto p(\mathcal{X} \mid \mathcal{S}, \mathcal{I}) p(\mathcal{S} \mid \mathcal{R}) \\ &= \left(\prod_i \prod_j \prod_n p(x_{i,j,n} \mid S_{I_i,n}) \right) \left(\prod_m \prod_n p(S_{m,n} \mid R_{m,n}) \right). \end{aligned}$$

Here, the likelihood of realigned read pairs can be rewritten as

$$\prod_i \prod_j \prod_n p(x_{i,j,n} \mid S_{I_i,n}) = \prod_m \prod_n \prod_{i:I_i=m} \prod_j p(x_{i,j,n} \mid S_{m,n}).$$

As a result, The conditional distribution of HLA sequences can be calculated by

$$p(\mathcal{S} \mid \mathcal{R}, \mathcal{I}, \mathcal{X}) \propto \prod_m \prod_n \left(\prod_{i:I_i=m} \prod_j p(x_{i,j,n} \mid S_{m,n}) \right) p(S_{m,n} \mid R_{m,n}).$$

Algorithm 4.1 Gibbs sampling of each HLA type in ALPHLARD

```
Input:
    $S_m$: the $m^{\mathrm{th}}$ HLA sequence
Output:
    $R_m$: the $m^{\mathrm{th}}$ HLA type

$T \leftarrow$ a set of HLA types
for all $t \in T$ do
    $p_t \leftarrow 1$
end for
for all $t \in T$ do
    for $n \leftarrow 1$ to $N$ do
        $p_t \leftarrow p_t \times p(S_{m,n} \mid R_{m,n}, R_m = t)$
    end for
    $p_t \leftarrow p_t \times p(R_m = t)$
end for
Sample $R_m$ with probability in proportion to $\boldsymbol{p}$
return $R_m$
```

Therefore, each HLA base $S_{m,n}$ can be sampled independently from the probability distribution in proportion to

$$\left(\prod_{i:I_i=m} \prod_j p(x_{i,j,n} \mid S_{m,n}) \right) p(S_{m,n} \mid R_{m,n}),$$

using Gibbs sampling. The Gibbs sampling of each HLA base is shown in Algorithm 4.2.

4.4.3 Gibbs Sampling of Indicator Variables $\mathcal{I}$

The conditional distribution of indicator variables is given by

$$\begin{aligned} p(\mathcal{I} \mid \mathcal{R}, \mathcal{S}, \mathcal{X}) &\propto p(\mathcal{X} \mid \mathcal{S}, \mathcal{I}) p(\mathcal{I}) \\ &= \prod_i \left(\prod_j \prod_n p(x_{i,j,n} \mid S_{I_i,n}) \right) p(I_i). \end{aligned}$$

Therefore, each indicator variable I_i can be sampled independently from the probability distribution in proportion to

$$\left(\prod_j \prod_n p(x_{i,j,n} \mid S_{I_i,n}) \right) p(I_i),$$

using Gibbs sampling. The Gibbs sampling of each indicator variable is shown in Algorithm 4.3.

4.4.4 Metropolis-Hastings Algorithm for HLA Bases Not Covered with Reads

In addition to Gibbs sampling, we use the Metropolis-Hastings algorithm with a proposal distribution that enables the parameters to jump from mode to mode and leads to more efficient sampling. This proposal distribution is focused on

Algorithm 4.2 Gibbs sampling of each HLA base in ALPHLARD

Input:

$\mathcal{R}$: HLA types
$\mathcal{I}$: indicator variables
$\mathcal{X}$: realigned read pairs

Output:

$\mathcal{S}$: HLA sequences

```
$K \leftarrow$ the number of realigned read pairs
$B \leftarrow \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}, \texttt{-}, \texttt{N}\}$
for $m \leftarrow 1$ to $\nu^{(\texttt{d})} + 2$ do
    for $n \leftarrow 1$ to $N$ do
        for $b \in B$ do
            $p_{m,n,b} \leftarrow 1$
        end for
    end for
end for
for $i \leftarrow 1$ to $K$ do
    $m \leftarrow I_i$
    $J \leftarrow \begin{cases} 1 & \text{(if } x_i \text{ is a paired read)} \\ 2 & \text{(if } x_i \text{ is an unpaired read)} \end{cases}$
    for $j \leftarrow 1$ to $J$ do
        $r \leftarrow$ a set of positions covered by the read $x_{i,j}$
        for $n \in r$ do
            for $b \in B$ do
                $p_{m,n,b} \leftarrow p_{m,n,b} \times p(x_{i,j,n} \mid S_{m,n} = b)$
            end for
        end for
    end for
end for
for $m \leftarrow 1$ to $\nu^{(\texttt{d})} + 2$ do
    for $n \leftarrow 1$ to $N$ do
        for $b \in B$ do
            $p_{m,n,b} \leftarrow p_{m,n,b} \times p(S_{m,n} = b \mid R_{m,n})$
        end for
    end for
end for
for $m \leftarrow 1$ to $\nu^{(\texttt{d})} + 2$ do
    for $n \leftarrow 1$ to $N$ do
        Sample $S_{m,n}$ with probability in proportion to $\boldsymbol{p_{m,n}}$
    end for
end for
return $\mathcal{S}$
```

Algorithm 4.3 Gibbs sampling of each indicator variable in ALPHLARD

```
Input:
    $\mathcal{S}$: HLA sequences
    $x_i$: the $i^{\text{th}}$ realigned read pair
Output:
    $I_i$: the $i^{\text{th}}$ indicator variable

$J \leftarrow \begin{cases} 1 & \text{(if } x_i \text{ is a paired read)} \\ 2 & \text{(if } x_i \text{ is an unpaired read)} \end{cases}$
for $m \leftarrow 1$ to $\nu^{(\mathtt{d})} + 2$ do
    $p_m \leftarrow 1$
end for
for $j \leftarrow 1$ to $J$ do
    $r \leftarrow$ a set of positions covered by the read $x_{i,j}$
    for $m \leftarrow 1$ to $\nu^{(\mathtt{d})} + 2$ do
        for $n \in r$ do
            $p_m \leftarrow p_m \times p(x_{i,j,n} \mid S_{m,n})$
        end for
    end for
end for
for $m \leftarrow 1$ to $\nu^{(\mathtt{d})} + 2$ do
    $p_m \leftarrow p_m \times p(I_i = m)$
end for
Sample $I_i$ with probability in proportion to $\boldsymbol{p}$
return $I_i$
```

positions not covered with any read. We first explain a problem when there is ambiguity in HLA types that is caused by some uncovered regions. For example, let t and t' be HLA types that have only one different base at the n^{th} position. If a sample has the HLA type t, but there are no reads that were produced by t and cover the n^{th} position, we cannot determine whether the sample has t or t'. In this case, once R_m becomes t', the next Gibbs sampling of $S_{m,n}$ gives $R_{m,n}$, or the n^{th} base of t', with high probability. Then, the next Gibbs sampling of R_m gives t' again with high probability. This process would be repeated, and it is difficult to move R_m from t' to t. This problem is caused by high correlation of the HLA type R_m and the HLA sequence S_m. To tackle the problem, we introduce the Metropolis-Hastings algorithm where the HLA type R_m and the HLA sequence S_m are simultaneously sampled. First, we define $S_m^{(\mathtt{N})}$ by

$$S_{m,n}^{(\mathtt{N})} = \begin{cases} S_{m,n} & \text{(if } \exists i;\ I_i = m \text{ and the } n^{\text{th}} \text{ base of } x_i \text{ is defined)} \\ \mathtt{N} & \text{(otherwise)} \end{cases} .$$

In other words, $S_m^{(\mathtt{N})}$ are basically the same as S_m, but bases that are not covered by any read are replaced with `N`'s. Then, A candidate HLA type R_m^* and a candidate HLA sequence S_m^* are sampled by

$$\begin{aligned} R_m^* &\sim p(R_m^* \mid S_m^{(\mathtt{N})}), \\ S_m^* &\sim p(S_m^* \mid R_m^*, \mathcal{S}_{-m}, \mathcal{I}, \mathcal{X}). \end{aligned}$$

Since $S_m^{(\mathtt{N})}$ have `N`'s at the positions where no reads cover, R_m^* can easily move among multiple types in the case that there is ambiguity mentioned above.

The acceptance ratio r is given by

$$\begin{aligned}
&r = \min(1, r^*),\\
&r^* = \frac{p(R_m^*, S_m^* \mid \mathcal{R}_{-m}, \mathcal{S}_{-m}, \mathcal{I}, \mathcal{X})}{p(R_m, S_m \mid \mathcal{R}_{-m}, \mathcal{S}_{-m}, \mathcal{I}, \mathcal{X})} \frac{p(R_m^*, S_m^* \to R_m, S_m \mid \mathcal{R}_{-m}, \mathcal{S}_{-m}, \mathcal{I}, \mathcal{X})}{p(R_m, S_m \to R_m^*, S_m^* \mid \mathcal{R}_{-m}, \mathcal{S}_{-m}, \mathcal{I}, \mathcal{X})}.
\end{aligned}$$

Each term can be obtained by

$$\begin{aligned}
&p(R_m, S_m \mid \mathcal{R}_{-m}, \mathcal{S}_{-m}, \mathcal{I}, \mathcal{X})\\
&\qquad \propto p(\mathcal{X} \mid S_m, \mathcal{S}_{-m}, \mathcal{I}) p(S_m \mid R_m) p(R_m),\\
&p(R_m, S_m \to R_m^*, S_m^* \mid \mathcal{R}_{-m}, \mathcal{S}_{-m}, \mathcal{I}, \mathcal{X})\\
&\qquad = p(R_m^* \mid S_m^{(\mathtt{N})}) p(S_m^* \mid R_m^*, \mathcal{I}, \mathcal{X})\\
&\qquad \propto p(S_m^{(\mathtt{N})} \mid R_m^*) p(R_m^*) \times \frac{p(\mathcal{X} \mid S_m^*, \mathcal{S}_{-m}, \mathcal{I}) p(S_m^* \mid R_m^*)}{p(\mathcal{X} \mid R_m^*, \mathcal{S}_{-m}, \mathcal{I})}.
\end{aligned}$$

As a result,

$$r^* = \frac{p(S_m^{(\mathtt{N})} \mid R_m)}{p(S_m^{(\mathtt{N})} \mid R_m^*)} \frac{p(\mathcal{X} \mid R_m^*, \mathcal{S}_{-m}, \mathcal{I})}{p(\mathcal{X} \mid R_m, \mathcal{S}_{-m}, \mathcal{I})}.$$

Here, $p(\mathcal{X} \mid R_m, \mathcal{S}_{-m}, \mathcal{I})$ can be calculated by

$$\begin{aligned}
&p(\mathcal{X} \mid R_m, \mathcal{S}_{-m}, \mathcal{I})\\
&\quad = \sum_S p(\mathcal{X} \mid S_m = S, \mathcal{S}, \mathcal{I}) p(S_m = S \mid R_m)\\
&\quad \propto \sum_S \left(\prod_{i:I_i = m} p(x_i \mid S_m = S) \right) p(S_m = S \mid R_m)\\
&\quad = \sum_S \left(\prod_{i:I_i = m} \prod_j \prod_n p(x_{i,j,n} \mid S_{m,n} = S_n) \right) \left(\prod_n p(S_{m,n} = S_n \mid R_{m,n}) \right)\\
&\quad = \sum_{b_1} \cdots \sum_{b_N} \prod_n \left(\prod_{i:I_i = m} \prod_j p(x_{i,j,n} \mid S_{m,n} = b_n) \right) p(S_{m,n} = b_n \mid R_{m,n})\\
&\quad = \prod_n \sum_{b_n} \left(\prod_{i:I_i = m} \prod_j p(x_{i,j,n} \mid S_{m,n} = b_n) \right) p(S_{m,n} = b_n \mid R_{m,n}).
\end{aligned}$$

4.4.5 Metropolis-Hastings Algorithm for Swapping Non-Decoy Parameters and Decoy Parameters

In addition, we use the Metropolis-Hastings algorithm with another proposal distribution that swaps non-decoy and decoy parameters, which helps to judge which HLA types and HLA sequences are non-decoy parameters. First, a non-decoy index m and a decoy index m' are uniformly sampled, that is,

$$\begin{aligned}
m &\sim \mathcal{U}(m \mid 1, 2),\\
m' &\sim \mathcal{U}(m' \mid 3, \nu^{(\mathtt{d})} + 2),
\end{aligned}$$

where $\mathcal{U}$ is a discrete uniform distribution. Then, the candidate HLA types R_m^* and $R_{m'}^*$ and the candidate HLA sequences S_m^* and $S_{m'}^*$ are obtained by swapping the m^{th} and m'^{th} HLA types and HLA sequences, that is,

$$\begin{aligned}
R^*_m &= R_{m'}, \\
R^*_{m'} &= R_m, \\
S^*_m &= S_{m'}, \\
S^*_{m'} &= S_m.
\end{aligned}$$

The acceptance ratio r is given by

$$\begin{aligned}
r &= \min(1, r^*), \\
r^* &= \frac{p(R^*_m, R^*_{m'}, S^*_m, S^*_{m'} \mid \mathcal{R}_{-m,m'}, \mathcal{S}_{-m,m'}, \mathcal{I}, \mathcal{X})}{p(R_m, R_{m'}, S_m, S_{m'} \mid \mathcal{R}_{-m,m'}, \mathcal{S}_{-m,m'}, \mathcal{I}, \mathcal{X})} \\
&\qquad \times \frac{p(R^*_m, R^*_{m'}, S^*_m, S^*_{m'} \to R_m, R_{m'}, S_m, S_{m'} \mid \mathcal{R}_{-m,m'}, \mathcal{S}_{-m,m'}, \mathcal{I}, \mathcal{X})}{p(R_m, R_{m'}, S_m, S_{m'} \to R^*_m, R^*_{m'}, S^*_m, S^*_{m'} \mid \mathcal{R}_{-m,m'}, \mathcal{S}_{-m,m'}, \mathcal{I}, \mathcal{X})},
\end{aligned}$$

where

$$\begin{aligned}
&p(R_m, R_{m'}, S_m, S_{m'} \mid \mathcal{R}_{-m,m'}, \mathcal{S}_{-m,m'}, \mathcal{I}, \mathcal{X}) \\
&\qquad \propto p(\mathcal{X} \mid S_m, S_{m'}, \mathcal{S}_{-m,m'}, \mathcal{I}) p(S_m \mid R_m) p(S_{m'} \mid R_{m'}) p(R_m), \\
&p(R_m, R_{m'}, S_m, S_{m'} \to R^*_m, R^*_{m'}, S^*_m, S^*_{m'} \mid \mathcal{R}_{-m,m'}, \mathcal{S}_{-m,m'}, \mathcal{I}, \mathcal{X}) \\
&\qquad \propto 1.
\end{aligned}$$

As a result, since

$$p(\mathcal{X} \mid S_m, S_{m'}, \mathcal{S}_{-m,m'}, \mathcal{I}) = p(\mathcal{X} \mid S^*_m, S^*_{m'}, \mathcal{S}_{-m,m'}, \mathcal{I}),$$

r^* is given by

$$r^* = \frac{p(S^*_m \mid R^*_m) p(S^*_{m'} \mid R^*_{m'}) p(R^*_m)}{p(S_m \mid R_m) p(S_{m'} \mid R_{m'}) p(R_m)}.$$

4.4.6 Strategies in the Burn-in Period

Other strategies were further used in the burn-in period to obtain better parameters. First, at the beginning of sampling, a multi-start strategy is used to obtain better initial parameters. Specifically, some MCMC chains are carried out, and initial parameters are sampled from the last parameters of the MCMC chains. Second, sequences of HLA types are sometimes copied to HLA sequences. This works well to get better parameters because there are many local optima where HLA sequences are twisted as if some crossovers occurred. Although this approach does not satisfy the detailed balance condition, which is a sufficient condition to obtain samples from the correct posterior distribution in MCMC, sampled parameters correctly reflect the posterior distribution since the approach is carried out only in the burn-in period.

4.4.7 HLA Analysis Using Sampled Parameters

After sampling the parameters, the posterior distribution of the HLA genotype can be calculated by counting sampled $R_1^{(\mathtt{r})}$ and $R_2^{(\mathtt{r})}$. The posterior distribution of the HLA sequences can be also inferred by counting $S_1^{(\mathtt{r})}$ and $S_2^{(\mathtt{r})}$. In the case that a sample has a novel HLA type, we can detect the fact from mismatches between HLA types and HLA sequences.

4.5 Experimental Results

In this section, we illustrate the capability of ALPHLARD using WES data and WGS data. First, we describe the detailed information on the used WES and WGS datasets. Then, we show the performance of ALPHLARD for HLA genotyping compared with other existing methods. In addition, we demonstrate that ALPHLARD can detect somatic mutations even from WGS data. Lastly, we discuss the effectiveness of decoy parameters in ALPHLARD.

In the following sections, we used the most sampled HLA genotype in the MCMC process as the candidate HLA genotype in ALPHLARD.

4.5.1 WES and WGS Datasets

To evaluate the capability of our method, we obtained 253 WES data with the HLA genotypes from the International HapMap Project [85] that had been used by Szolek *et al.* [81] and Shukla *et al.* [78]. We further downsampled these data to 1/2, 1/4, 1/8, and 1/16 to simulate low-coverage data.

We also used paired normal and tumor WGS data of 25 Japanese cancer patients, including 20 liver cancer and 5 microsatellite-unstable colon cancer samples. These data were obtained from an Illumina HiSeq system with a 101-bp pair-end read length.

The sequence-based typing (SBT) approach, which is guaranteed to be accurate at second field resolution, was used for validation of the 20 liver cancer samples. Additional HLA genotyping using the TruSight HLA Sequencing Panels [90], which are theoretically guaranteed to be accurate at full resolution, was performed for 7 out of the above 20 liver cancer samples to reduce ambiguity of the SBT genotyping. The 5 microsatellite-unstable samples were genotyped using the TruSight HLA Sequencing Panels, in order to verify not only the HLA genotypes but also the presence of somatic mutations. We regarded the results of the SBT approach and/or the TruSight HLA Sequencing Panels as the correct information. If the results differed between the two methods, we assumed that the result of the TruSight HLA Sequencing Panel was correct.

4.5.2 WES-based and WGS-based HLA Genotyping

For performance comparison, we used three existing methods, OptiType [81], PHLAT [3], and HLA-VBSeq [62] because it has been reported that they achieve the highest accuracy for WES-based and WGS-based HLA genotyping [4]. First, we applied ALPHLARD and the existing methods to the original and the downsampled WES data. Because the gold standard HLA genotypes were determined from exon 2 and 3, we used only the exons as the reference sequences in ALPHLARD. Figure 4.2 shows the performance of the methods. ALPHLARD keeps higher accuracy compared with the other methods even when the downsampling ratio is low. The accuracy of the existing methods seems consistent with the preceding paper [4].

We also applied the methods to the normal WGS data and compared the determined HLA genotypes with those obtained by the SBT approach and the TruSight HLA Sequencing Panel. The performance of the four methods is shown in Tables 4.1 and 4.2. Table 4.1 shows how many HLA types were correctly determined, and Table 4.2 shows how many samples were fully correctly genotyped. The tables demonstrate that ALPHLARD achieved a higher accuracy rate than the other methods for all of the HLA genes at any resolution. In addition, ALPHLARD correctly identified an HLA-B type in a sample, which was

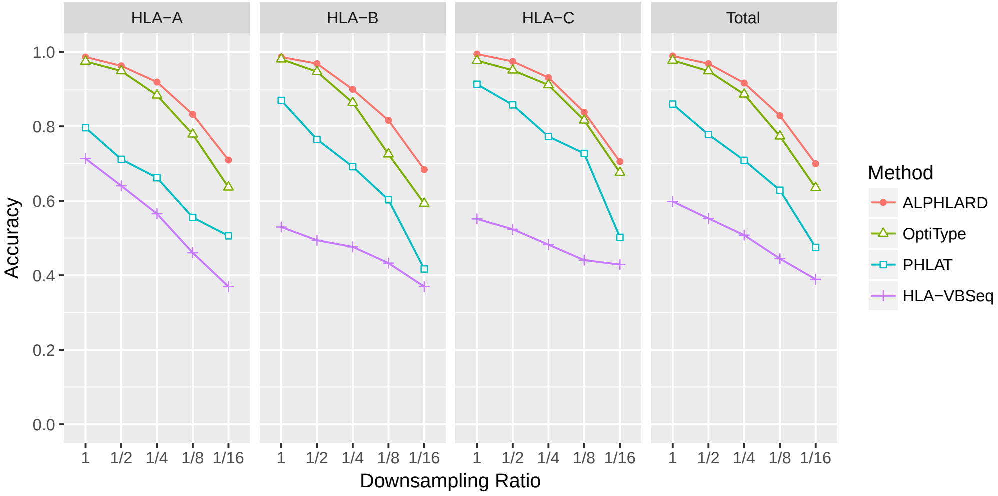


Figure 4.2: WES-based HLA genotyping of ALPHLARD, OptiType, PHLAT, and HLA-VBSeq. Each WES data was downsampled to 1/2, 1/4, 1/8, and 1/16, and the four methods were applied to all of the original and the downsampled WES data.

determined differently by the SBT approach and the TruSight HLA Sequencing Panel. This suggests that ALPHLARD could be potentially superior to the SBT approach in some cases. OptiType achieved the best performance for HLA class I genotyping among the existing methods. Also, The results of WES-based and WGS-based HLA genotyping show that HLA-VBSeq achieved relatively high accuracy for the WGS data compared with the WES data, which means that HLA-VBSeq would take advantage of sequence reads from non-coding regions such as the introns and the untranslated regions. The accuracy of the existing methods was consistent with the preceding paper [4].

Table 4.1: WGS-based HLA genotyping accuracy that indicates how many HLA types were correctly determined with ALPHLARD, OptiType, PHLAT, and HLA-VBSeq. N/A indicates that the method does not support the HLA gene or the resolution.

		ALPHLARD	OptiType	PHLAT	HLA-VBSeq
HLA-A	1st	**100% (50/50)**	**100% (50/50)**	76.0% (38/50)	96.0% (48/50)
	2nd	**98.0% (49/50)**	**98.0% (49/50)**	60.0% (30/50)	82.0% (41/50)
	3rd	**98.0% (49/50)**	N/A	46.0% (23/50)	82.0% (41/50)
HLA-B	1st	**100% (48/48)**	87.5% (42/48)	72.9% (35/48)	89.6% (43/48)
	2nd	**100% (48/48)**	85.4% (41/48)	56.3% (27/48)	75.0% (36/48)
	3rd	**95.8% (46/48)**	N/A	39.6% (19/48)	72.9% (35/48)
HLA-C	1st	**100% (50/50)**	**100% (50/50)**	78.0% (39/50)	96.0% (48/50)
	2nd	**98.0% (49/50)**	94.0% (47/50)	56.0% (28/50)	66.0% (33/50)
	3rd	**98.0% (49/50)**	N/A	44.0% (22/50)	66.0% (33/50)
HLA-DPA1	1st	**100% (24/24)**	N/A	N/A	87.5% (21/24)
	2nd	**100% (24/24)**	N/A	N/A	87.5% (21/24)
	3rd	**100% (24/24)**	N/A	N/A	87.5% (21/24)
HLA-DPB1	1st	**100% (22/22)**	N/A	N/A	86.4% (19/22)
	2nd	**100% (22/22)**	N/A	N/A	86.4% (19/22)
	3rd	**100% (22/22)**	N/A	N/A	86.4% (19/22)
HLA-DQA1	1st	**100% (24/24)**	N/A	70.8% (17/24)	**100% (24/24)**
	2nd	**95.8% (23/24)**	N/A	62.5% (15/24)	**95.8% (23/24)**
	3rd	**95.8% (23/24)**	N/A	62.5% (15/24)	**95.8% (23/24)**
HLA-DQB1	1st	**100% (18/18)**	N/A	77.8% (14/18)	**100% (18/18)**
	2nd	**94.4% (17/18)**	N/A	61.1% (11/18)	88.9% (16/18)
	3rd	**94.4% (17/18)**	N/A	38.9% (7/18)	88.9% (16/18)
HLA-DRB1	1st	**100% (24/24)**	N/A	70.8% (17/24)	95.8% (23/24)
	2nd	**100% (24/24)**	N/A	50.0% (12/24)	58.3% (14/24)
	3rd	**100% (24/24)**	N/A	45.8% (11/24)	58.3% (14/24)
Total	1st	**100% (260/260)**	95.9% (142/148)	74.8% (160/214)	93.8% (244/260)
	2nd	**98.5% (256/260)**	92.6% (137/148)	57.5% (123/214)	78.1% (203/260)
	3rd	**97.7% (254/260)**	N/A	45.3% (97/214)	77.7% (202/260)

Table 4.2: WGS-based HLA genotyping accuracy that indicates how many samples were fully correctly genotyped with ALPHLARD, OptiType, PHLAT, and HLA-VBSeq. N/A indicates that the method does not support the HLA gene or the resolution.

		ALPHLARD	OptiType	PHLAT	HLA-VBSeq
HLA-A	1st	**100% (25/25)**	**100% (25/25)**	64.0% (16/25)	92.0% (23/25)
	2nd	**96.0% (24/25)**	**96.0% (24/25)**	36.0% (9/25)	72.0% (18/25)
	3rd	**96.0% (24/25)**	N/A	20.0% (5/25)	72.0% (18/25)
HLA-B	1st	**100% (24/24)**	79.2% (19/24)	66.7% (16/24)	79.2% (19/24)
	2nd	**100% (24/24)**	75.0% (18/24)	45.8% (11/24)	54.2% (13/24)
	3rd	**91.7% (22/24)**	N/A	25.0% (6/24)	50.0% (12/24)
HLA-C	1st	**100% (25/25)**	**100% (25/25)**	76.0% (19/25)	92.0% (23/25)
	2nd	**96.0% (24/25)**	92.0% (23/25)	44.0% (11/25)	40.0% (10/25)
	3rd	**96.0% (24/25)**	N/A	28.0% (7/25)	40.0% (10/25)
HLA-DPA1	1st	**100% (12/12)**	N/A	N/A	75.0% (9/12)
	2nd	**100% (12/12)**	N/A	N/A	75.0% (9/12)
	3rd	**100% (12/12)**	N/A	N/A	75.0% (9/12)
HLA-DPB1	1st	**100% (11/11)**	N/A	N/A	81.8% (9/11)
	2nd	**100% (11/11)**	N/A	N/A	81.8% (9/11)
	3rd	**100% (11/11)**	N/A	N/A	81.8% (9/11)
HLA-DQA1	1st	**100% (12/12)**	N/A	66.7% (8/12)	**100% (12/12)**
	2nd	**91.7% (11/12)**	N/A	50.0% (6/12)	**91.7% (11/12)**
	3rd	**91.7% (11/12)**	N/A	50.0% (6/12)	**91.7% (11/12)**
HLA-DQB1	1st	**100% (9/9)**	N/A	77.8% (7/9)	**100% (9/9)**
	2nd	**88.9% (8/9)**	N/A	44.4% (4/9)	77.8% (7/9)
	3rd	**88.9% (8/9)**	N/A	33.3% (3/9)	77.8% (7/9)
HLA-DRB1	1st	**100% (12/12)**	N/A	58.3% (7/12)	91.7% (11/12)
	2nd	**100% (12/12)**	N/A	33.3% (4/12)	41.7% (5/12)
	3rd	**100% (12/12)**	N/A	33.3% (4/12)	41.7% (5/12)
Total	1st	**100% (130/130)**	93.2% (69/74)	68.2% (73/107)	88.5% (115/130)
	2nd	**96.9% (126/130)**	87.8% (65/74)	42.1% (45/107)	63.1% (82/130)
	3rd	**95.4% (124/130)**	N/A	29.0% (31/107)	62.3% (81/130)

4.5.3 Detection of Somatic Mutations

Next, we searched for somatic point mutations in the HLA genes. They can be detected by comparing the inferred HLA sequences between paired normal and tumor samples of each patient. We detected three somatic point mutations in the microsatellite-unstable samples: two single-base deletions and one single-base insertion (Figures 4.3–4.5). One of the deletions occurred in a homopolymeric region in exon 1 of the HLA-A gene, and the other occurred in a homopolymeric region in exon 1 of the HLA-B gene. Both of these mutations caused a frameshift, leading to an early stop codon and ultimate loss of function of the HLA allele. It is known that the HLA-A and HLA-B genes are paralogous, and we found that the two deletions occurred at paralogously the same position. Moreover, an HLA-A type A*68:11N has a single-base deletion at exactly the same homopolymeric position. These observations suggest that the homopolymeric region is a deletion hotspot. Also, the insertion occurred in a homopolymeric region at the beginning of exon 4 of the HLA-A gene, which changed the HLA-A type from A*31:01:02 to A*31:14N. This region is known as an insertion hotspot in some HLA types such as A*01:04N and B*51:11N, and the insertion causes no expression of the allele [41, 53, 79, 14]. The three insertions and deletions identified were validated by the TruSight HLA Sequencing Panels.

We further sought cases of loss of heterozygosity (LOH), which is a genetic event that causes loss of a region in a chromosome, in the HLA genes as follows. First, we focused on two types of patients: (i) those for which HLA genotypes were uniquely determined for the normal sample but not for the tumor sample, and (ii) those for which HLA genotypes of both the normal and the tumor samples were uniquely but not identically determined. Then, we checked whether the collected reads of the tumor sample supported the HLA genotype inferred for the normal sample. We were able to detect one likely case of LOH in the tumor sample of a patient, RK069. At each heterozygous SNP position in each HLA locus, the log odds ratio was calculated for the WGS data and the TruSight HLA Sequencing Panels based on the number of reads that supported the SNP (Figures 4.6–4.11). These figures suggest that A*26:01:01, B*35:01:01, C*03:03:01, DPA1*01:03:01, DQA1*03:02, and DRB1*12:01:01 might be lost in the tumor sample of RK069.

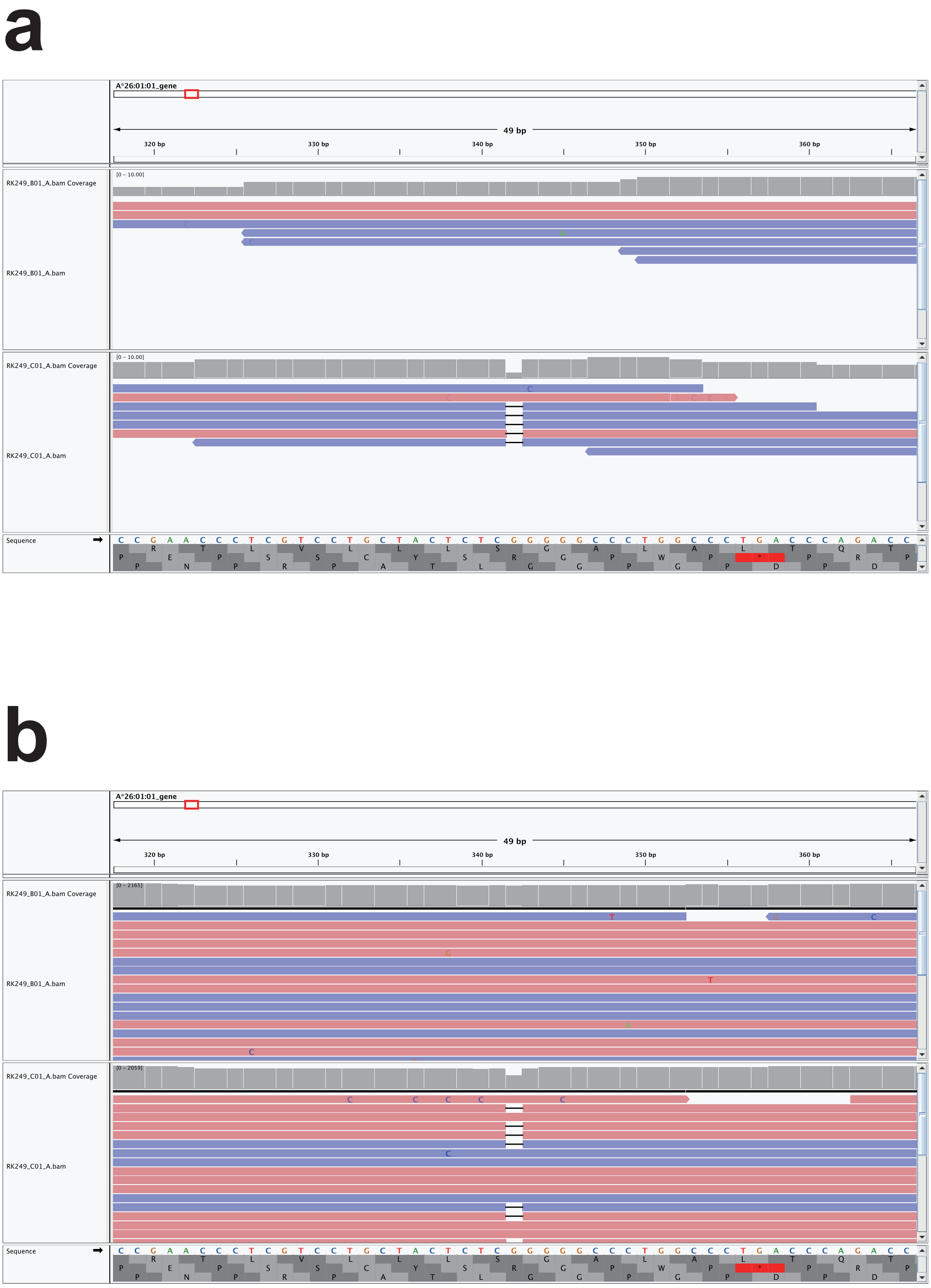


Figure 4.3: A single-base deletion in exon 1 of the HLA-A gene of patient RK249. IGV screenshots were taken at the position for (a) the WGS data and (b) the TruSight HLA Sequencing Panel data. In each of the screenshots, the upper and lower tracks correspond to the normal and tumor samples, respectively.

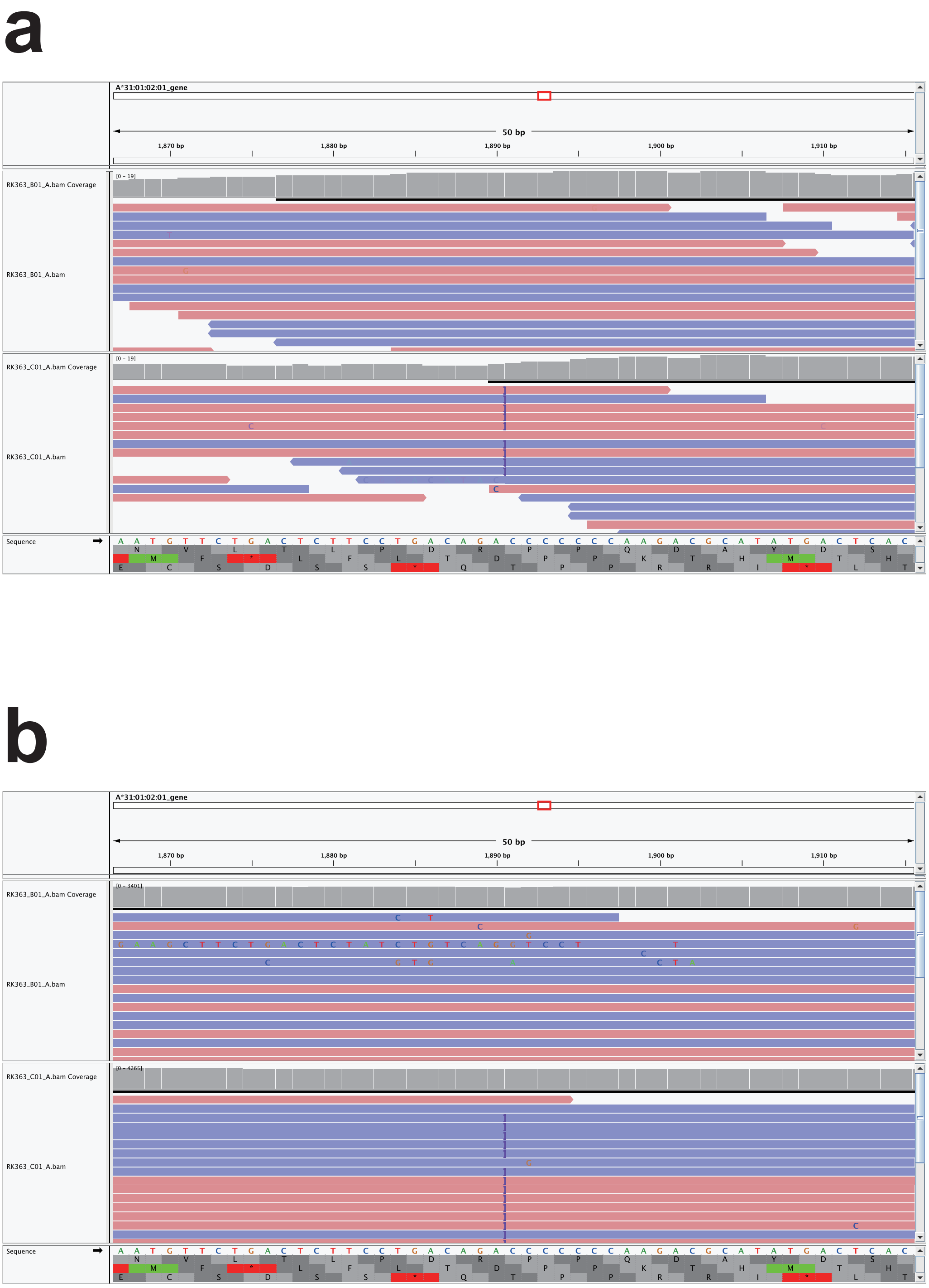


Figure 4.4: A single-base insertion in exon 4 of the HLA-A gene of patient RK363. IGV screenshots were taken at the position for (a) the WGS data and (b) the TruSight HLA Sequencing Panel data. In each of the screenshots, the upper and lower tracks correspond to the normal and tumor samples, respectively.

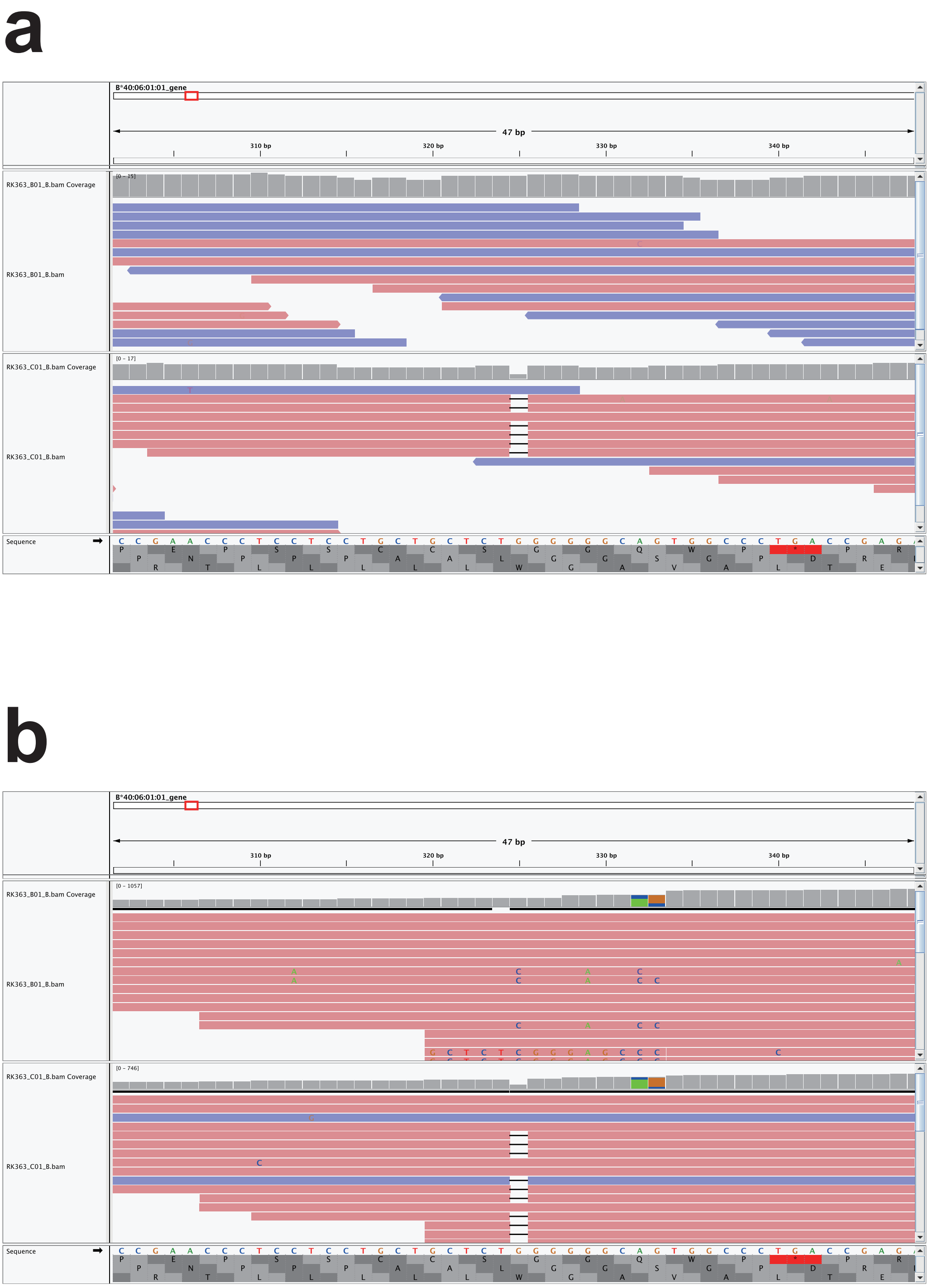


Figure 4.5: A single-base deletion in exon 1 of the HLA-B gene of patient RK363. IGV screenshots were taken at the position for (a) the WGS data and (b) the TruSight HLA Sequencing Panel data. In each of the screenshots, the upper and lower tracks correspond to the normal and tumor samples, respectively.

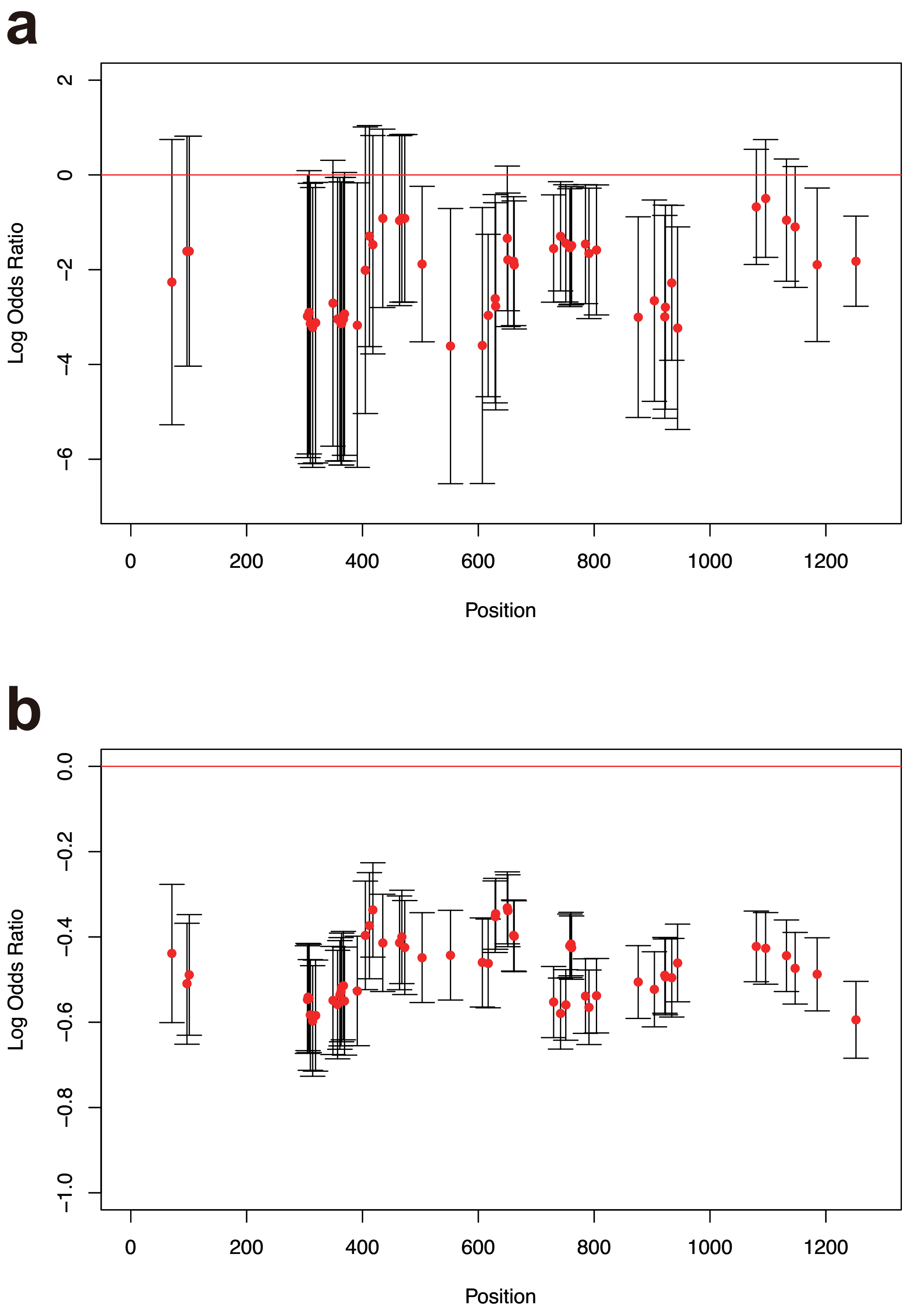


Figure 4.6: The log odds ratios of the depths at heterozygous SNP positions in the HLA-A gene of patient RK069. The log odds ratios were calculated for (a) the WGS data and (b) the TruSight HLA Sequencing Panel data. These log odds ratios correspond to the relative quantities of observed A*26:01:01 SNPs in the tumor sample compared with the normal sample. The red dots indicate the mean values of the log odds ratios, and the vertical lines indicate the 95% confidence intervals.

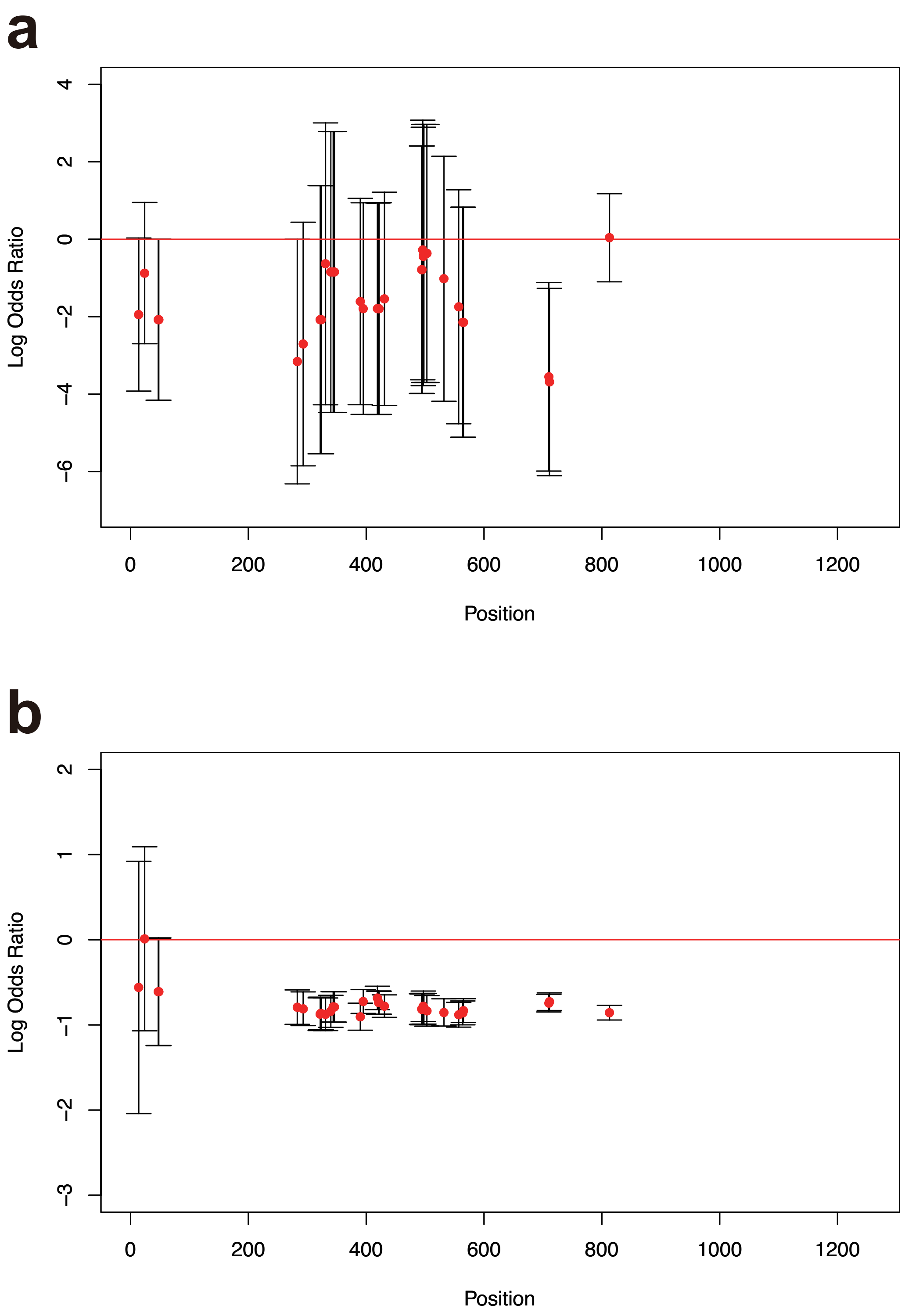


Figure 4.7: The log odds ratios of the depths at heterozygous SNP positions in the HLA-B gene of patient RK069. The log odds ratios were calculated for (a) the WGS data and (b) the TruSight HLA Sequencing Panel data. These log odds ratios correspond to the relative quantities of observed B*35:01:01 SNPs in the tumor sample compared with the normal sample. The red dots indicate the mean values of the log odds ratios, and the vertical lines indicate the 95% confidence intervals.

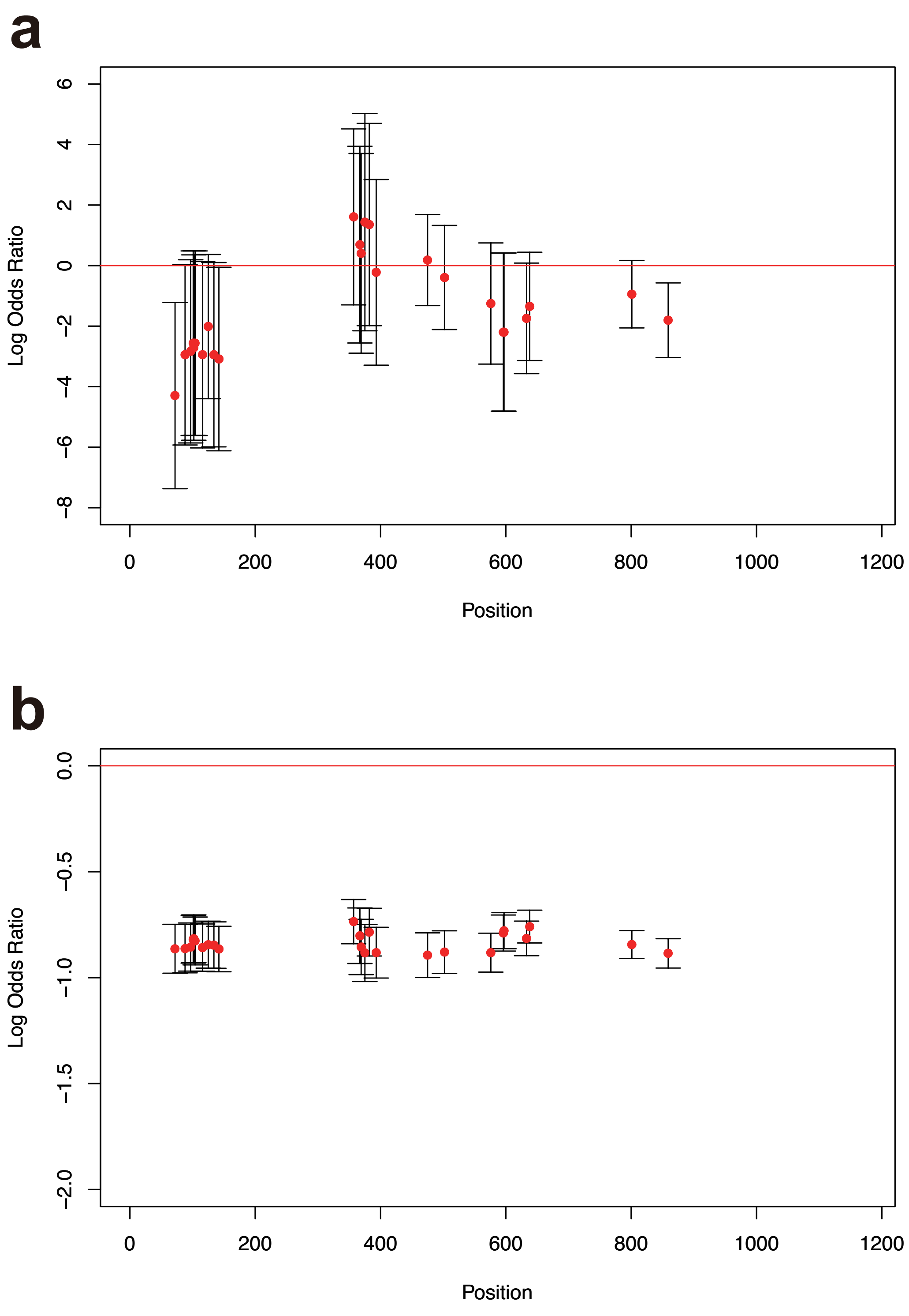


Figure 4.8: The log odds ratios of the depths at heterozygous SNP positions in the HLA-C gene of patient RK069. The log odds ratios were calculated for (a) the WGS data and (b) the TruSight HLA Sequencing Panel data. These log odds ratios correspond to the relative quantities of observed C*03:03:01 SNPs in the tumor sample compared with the normal sample. The red dots indicate the mean values of the log odds ratios, and the vertical lines indicate the 95% confidence intervals.

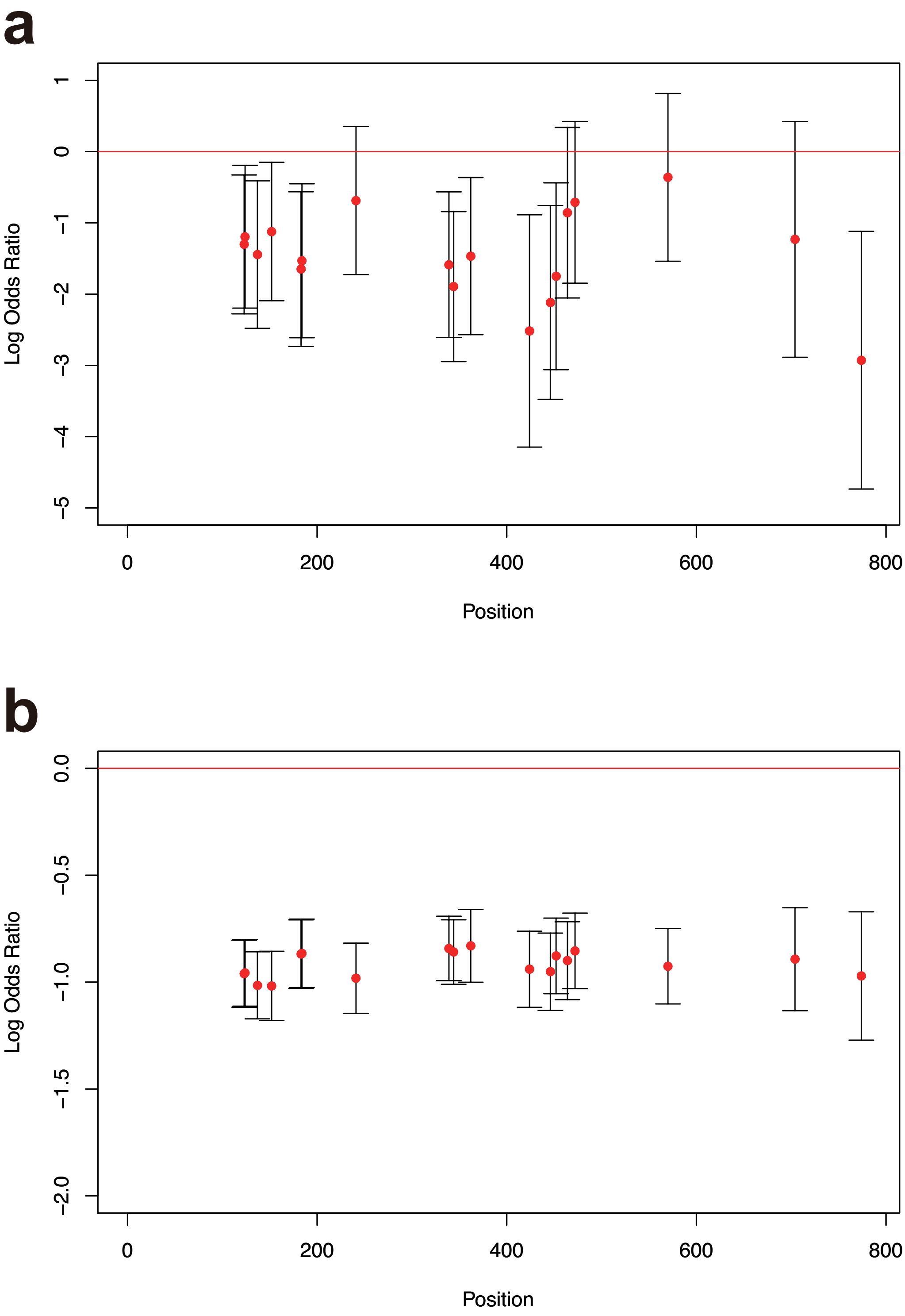


Figure 4.9: The log odds ratios of the depths at heterozygous SNP positions in the HLA-DPA1 gene of patient RK069. The log odds ratios were calculated for (a) the WGS data and (b) the TruSight HLA Sequencing Panel data. These log odds ratios correspond to the relative quantities of observed DPA1*01:03:01 SNPs in the tumor sample compared with the normal sample. The red dots indicate the mean values of the log odds ratios, and the vertical lines indicate the 95% confidence intervals.

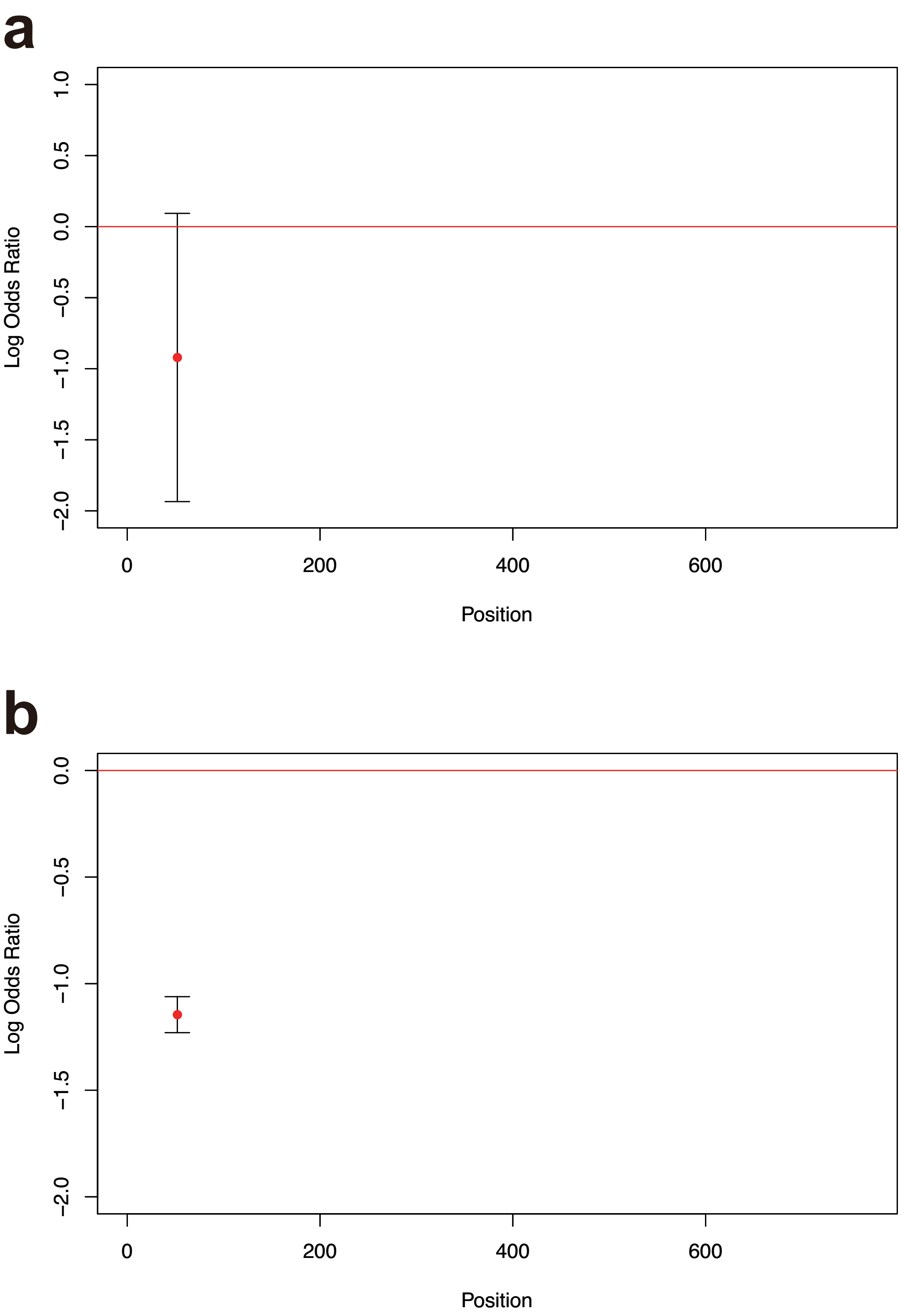


Figure 4.10: The log odds ratios of the depths at heterozygous SNP positions in the HLA-DQA1 gene of patient RK069. The log odds ratios were calculated for (a) the WGS data and (b) the TruSight HLA Sequencing Panel data. These log odds ratios correspond to the relative quantities of observed DQA1*03:02 SNPs in the tumor sample compared with the normal sample. The red dots indicate the mean values of the log odds ratios, and the vertical lines indicate the 95% confidence intervals.

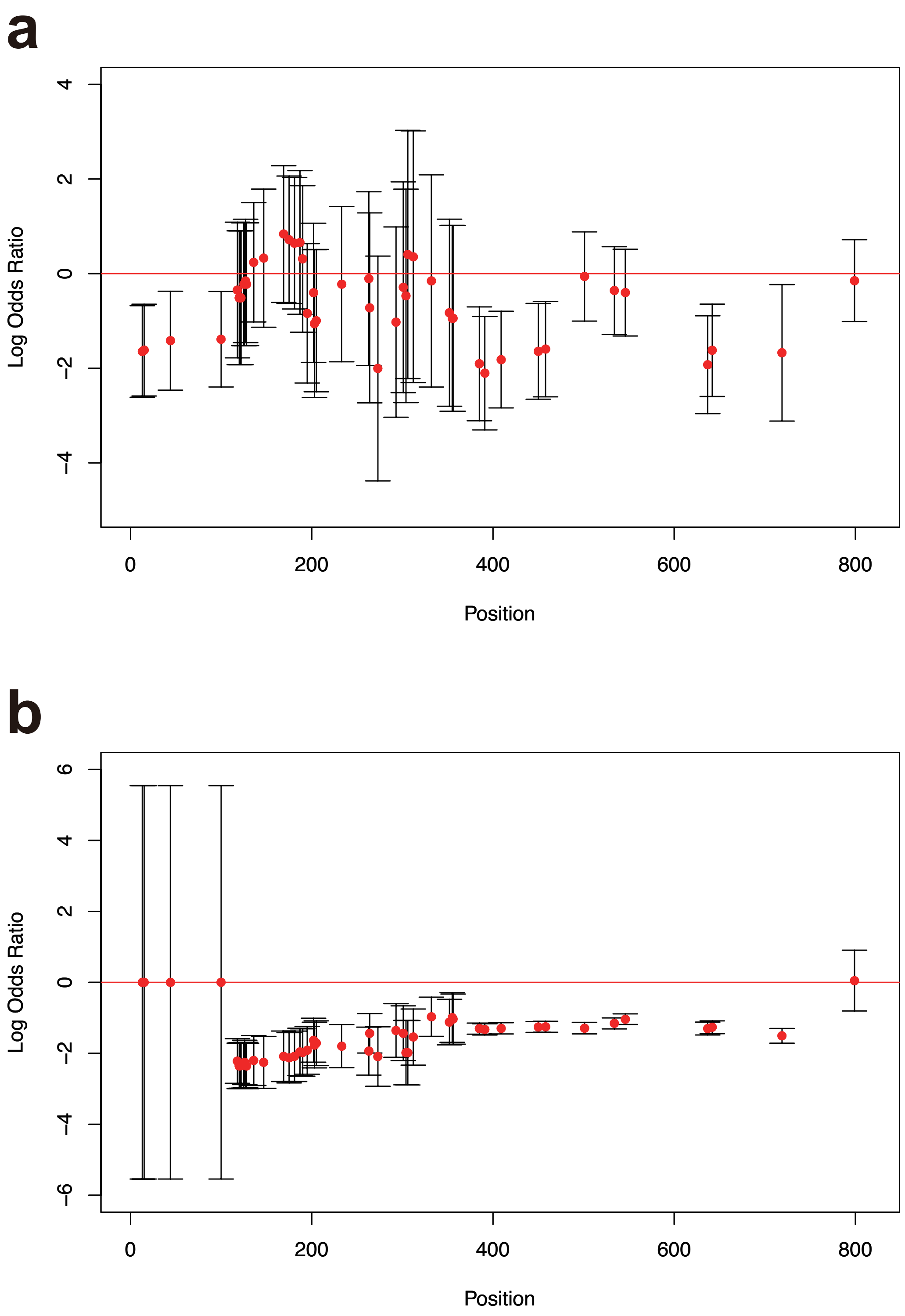


Figure 4.11: The log odds ratios of the depths at heterozygous SNP positions in the HLA-DRB1 gene of patient RK069. The log odds ratios were calculated for (a) the WGS data and (b) the TruSight HLA Sequencing Panel data. These log odds ratios correspond to the relative quantities of observed DRB1*12:01:01 SNPs in the tumor sample compared with the normal sample. The red dots indicate the mean values of the log odds ratios, and the vertical lines indicate the 95% confidence intervals.

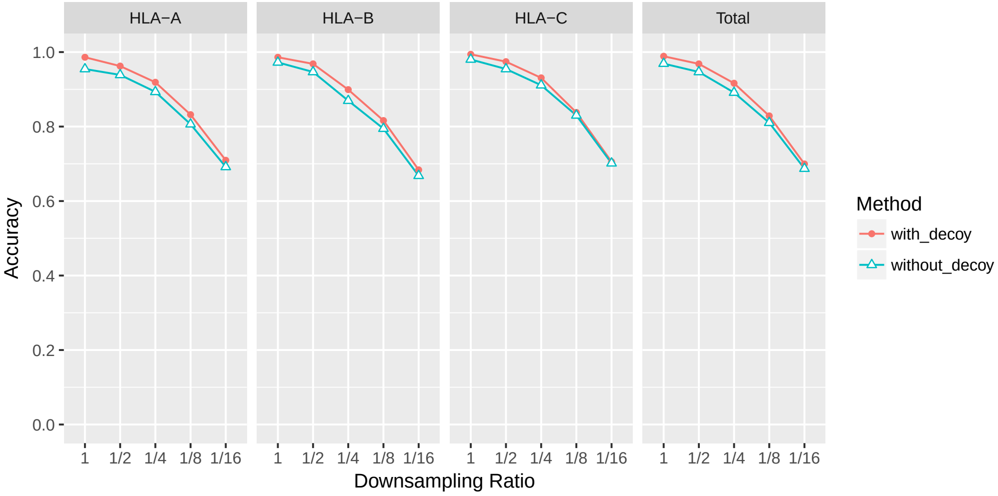


Figure 4.12: WES-based HLA genotyping accuracy of ALPHLARD with decoy parameters and without decoy parameters. Each WES data was downsampled to 1/2, 1/4, 1/8, and 1/16, and the two approaches were applied to all of the original and the downsampled WES data.

4.5.4 Effectiveness of Decoy Parameters

We next demonstrate the effectiveness of decoy parameters in ALPHLARD. First, we applied two versions of ALPHLARD, ALPHLARD with decoy parameters and ALPHLARD without decoy parameters, to the WES data used in the previous section. Figure 4.12 shows the performance of the two versions. At any HLA locus and any downsampling ratio, ALPHLARD with decoy parameters outperformed ALPHLARD without decoy parameters. This suggests that decoy parameters would reduce the influence of misclassified reads from other HLA genes and HLA pseudogenes. Also, the difference in the performance is significant when the downsampling ratio is high. This implies that the influence of misclassified reads cannot be ignorant when the sequence data contains a lot of reads.

We also applied the two approaches to the normal WGS data used above. Tables 4.3 and 4.4 show the performance of the approaches. These tables demonstrate that the performance of ALPHLARD with decoy parameters is higher than that of ALPHLARD without decoy parameters also for the WGS data.

Table 4.3: WGS-based HLA genotyping accuracy that indicates how many HLA types were correctly determined by ALPHLARD with decoy parameters and without decoy parameters.

		with decoy	without decoy
HLA-A	1st	**100% (50/50)**	**100% (50/50)**
	2nd	**98.0% (49/50)**	94.0% (47/50)
	3rd	**98.0% (49/50)**	94.0% (47/50)
HLA-B	1st	**100% (48/48)**	**100% (48/48)**
	2nd	**100% (48/48)**	97.9% (47/48)
	3rd	**95.8% (46/48)**	93.8% (45/48)
HLA-C	1st	**100% (50/50)**	**100% (50/50)**
	2nd	**98.0% (49/50)**	**98.0% (49/50)**
	3rd	**98.0% (49/50)**	**98.0% (49/50)**
HLA-DPA1	1st	**100% (24/24)**	**100% (24/24)**
	2nd	**100% (24/24)**	**100% (24/24)**
	3rd	**100% (24/24)**	**100% (24/24)**
HLA-DPB1	1st	**100% (22/22)**	**100% (22/22)**
	2nd	**100% (22/22)**	**100% (22/22)**
	3rd	**100% (22/22)**	**100% (22/22)**
HLA-DQA1	1st	**100% (24/24)**	**100% (24/24)**
	2nd	**95.8% (23/24)**	91.7% (22/24)
	3rd	**95.8% (23/24)**	91.7% (22/24)
HLA-DQB1	1st	**100% (18/18)**	88.9% (16/18)
	2nd	**94.4% (17/18)**	72.2% (13/18)
	3rd	**94.4% (17/18)**	72.2% (13/18)
HLA-DRB1	1st	**100% (24/24)**	**100% (24/24)**
	2nd	**100% (24/24)**	95.8% (23/24)
	3rd	**100% (24/24)**	95.8% (23/24)
Total	1st	**100% (260/260)**	99.2% (258/260)
	2nd	**98.5% (256/260)**	95.0% (247/260)
	3rd	**97.7% (254/260)**	94.2% (245/260)

Table 4.4: WGS-based HLA genotyping accuracy that indicates how many samples were fully correctly genotyped by ALPHLARD with decoy parameters and without decoy parameters.

		with decoy	without decoy
HLA-A	1st	**100% (25/25)**	**100% (25/25)**
	2nd	**96.0% (24/25)**	88.0% (22/25)
	3rd	**96.0% (24/25)**	88.0% (22/25)
HLA-B	1st	**100% (24/24)**	**100% (24/24)**
	2nd	**100% (24/24)**	95.8% (23/24)
	3rd	**91.7% (22/24)**	87.5% (21/24)
HLA-C	1st	**100% (25/25)**	**100% (25/25)**
	2nd	**96.0% (24/25)**	**96.0% (24/25)**
	3rd	**96.0% (24/25)**	**96.0% (24/25)**
HLA-DPA1	1st	**100% (12/12)**	**100% (12/12)**
	2nd	**100% (12/12)**	**100% (12/12)**
	3rd	**100% (12/12)**	**100% (12/12)**
HLA-DPB1	1st	**100% (11/11)**	**100% (11/11)**
	2nd	**100% (11/11)**	**100% (11/11)**
	3rd	**100% (11/11)**	**100% (11/11)**
HLA-DQA1	1st	**100% (12/12)**	**100% (12/12)**
	2nd	**91.7% (11/12)**	83.3% (10/12)
	3rd	**91.7% (11/12)**	83.3% (10/12)
HLA-DQB1	1st	**100% (9/9)**	77.8% (7/9)
	2nd	**88.9% (8/9)**	55.6% (5/9)
	3rd	**88.9% (8/9)**	55.6% (5/9)
HLA-DRB1	1st	**100% (12/12)**	**100% (12/12)**
	2nd	**100% (12/12)**	91.7% (11/12)
	3rd	**100% (12/12)**	91.7% (11/12)
Total	1st	**100% (130/130)**	98.5% (128/130)
	2nd	**96.9% (126/130)**	90.8% (118/130)
	3rd	**95.4% (124/130)**	89.2% (116/130)

Chapter 5

Bayesian Approach for HLA Somatic Mutation Calling from Whole Genome Sequence Data

5.1 Overview

Recently, the interaction between cancer and the immune system has attracted attention. Recent studies have shown that somatic mutations in HLA genes tend to accumulate in specific cancer types. Since these somatic mutations can contribute to suppressing the ability of the immune system and developing the tumor, it is considered to be important to accurately call HLA somatic mutations. However, Identification of HLA somatic mutations is generally difficult because true HLA mutations must be distinguished from false-positive mutations caused by similarity in HLA genes and HLA pseudogenes. This is true of ALPHLARD, which is our method introduced in Chapter 4.

To resolve this issue, we extend ALPHLARD, which was described in Chapter 4, to construct a new Bayesian model, named ALPHLARD-NT, for accurate HLA analysis from WGS data including both HLA germline and somatic mutation calling as well as HLA genotyping. ALPHLARD-NT conducts HLA analysis for each HLA locus independently by using both normal and tumor reads that were classified into the HLA locus in Chapter 3.3. ALPHLARD-NT infers HLA genotypes, HLA germline mutations, and HLA somatic mutations from sampled parameters in MCMC.

The organization of this chapter is as follows. First, in Section 5.2, we introduce an existing method, POLYSOLVER [78], that is used for performance comparison. Then, in Section 5.3, we explain how ALPHLARD-NT is constructed in detail. In Section 5.4, we explain how parameters are sampled from the Bayesian model. Lastly, we show some experimental results in 5.5.

The results in this chapter have been accepted as reference [27].

5.2 Related Work

5.2.1 POLYSOLVER

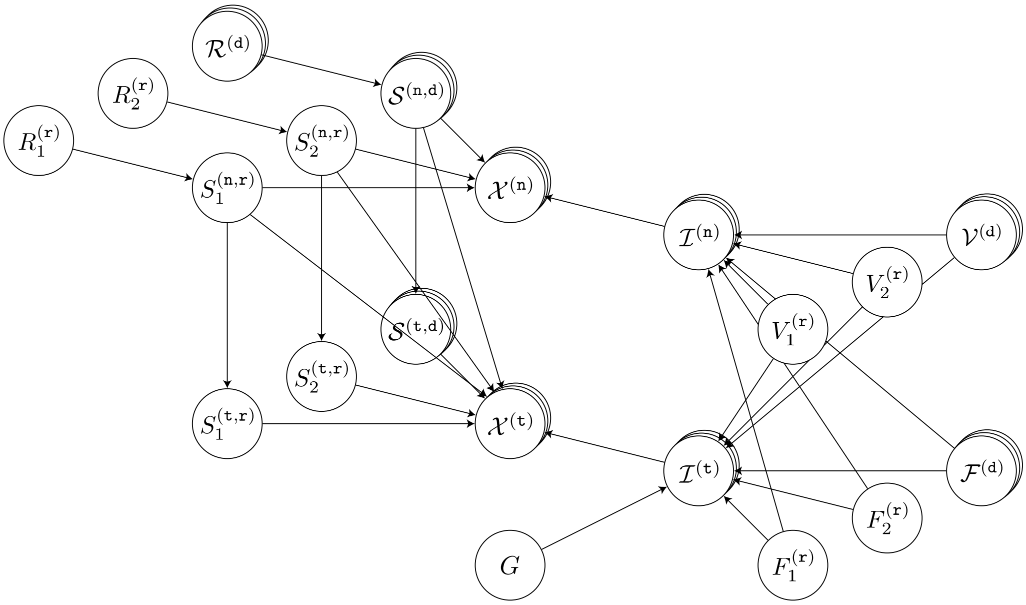
POLYSOLVER [78] is a method that can identify HLA somatic mutations as well as HLA genotypes at second field resolution from paired-ended WES data. POLYSOLVER performs HLA genotyping and HLA somatic mutation calling for the HLA-A, HLA-B, and HLA-C genes. First, POLYSOLVER determines HLA genotypes using a Bayesian approach. As in the case of our methods, the Allele Frequency Net Database [21] is used to calculate prior probabilities of HLA types. One of the significant characteristics of POLYSOLVER is that it can take as input the ethnicity information of samples. The ethnicity information is helpful for HLA genotyping because distributions of HLA types are different among human populations [10]. For each HLA type, in addition to the prior probability, the likelihood of reads given the HLA type is also calculated. Based on the prior probability and the likelihood, the posterior probability of each HLA type given reads can be obtained. POLYSOLVER employs the HLA type that achieves the highest posterior probability as the first HLA type of the sample. The second HLA type is also chosen in a similar way, except that the likelihoods are weighted using the likelihoods for the first HLA type to prefer HLA type. This weighting is designed to prefer HLA types that are not similar to the first HLA type. Consequently, the second HLA type is chosen in such a way that reads are exclusively aligned to the first and the second HLA types.

After HLA genotyping, POLYSOLVER realigns reads to the identified HLA genotypes. Then, POLYSOLVER conducts HLA somatic mutation calling using MuTect [9] and Strelka [74], which are standard mutation callers that are not limited to HLA genes. MuTect is used to detect HLA somatic substitutions. On the other hand, Strelka is used to identify insertions and deletions.

5.3 Bayesian Model for HLA Analysis

ALPHLARD-NT has partially the same structure as ALPHLARD except for some additional parameters. Figure 5.1 shows the graphical model. Input data of the model include both the normal and tumor realigned reads. Let $x_i^{(\mathtt{n})}$ be the i^{th} normal realigned paired reads or unpaired read, and $x_i^{(\mathtt{t})}$ be the i^{th} tumor realigned paired reads or unpaired read, both of which are obtained in Chapter 3. Here, $\mathtt{n}$ and $\mathtt{t}$ indicate parameters for the normal and tumor samples, respectively. We define $x_{i,j,n}^{(\mathtt{n})}$ and $x_{i,j,n}^{(\mathtt{t})}$ as the n^{th} bases or gaps of $x_{i,j}^{(\mathtt{n})}$ and $x_{i,j}^{(\mathtt{t})}$, respectively, and $p_{i,j,n}^{(\mathtt{n})}$ and $p_{i,j,n}^{(\mathtt{t})}$ as the mismatch probabilities of $x_{i,j,n}^{(\mathtt{n})}$ and $x_{i,j,n}^{(\mathtt{t})}$, respectively. Note that the first position of each realigned read is not the beginning of the read but rather the beginning of the combined MSA. Also, let $r_{i,j}^{(\mathtt{n})}$ and $r_{i,j}^{(\mathtt{t})}$ be sets of positions covered by the reads $x_{i,j}^{(\mathtt{n})}$ and $x_{i,j}^{(\mathtt{t})}$, respectively.

We denote HLA types of the sample by $R_1^{(\mathrm{r})}$ and $R_2^{(\mathrm{r})}$, normal HLA sequences by $S_1^{(\mathrm{n,r})}$ and $S_2^{(\mathrm{n,r})}$. In addition to normal HLA sequences, we introduce new parameters for tumor HLA sequences $S_1^{(\mathtt{t,r})}$ and $S_2^{(\mathtt{t,r})}$, which are used to consider somatic mutations in the tumor sample. We also introduce decoy HLA types $R_1^{(\mathtt{d})}, \ldots, R_{\nu^{(\mathtt{d})}}^{(\mathtt{d})}$, decoy normal HLA sequences $S_1^{(\mathtt{n,d})}, \ldots, S_{\nu^{(\mathtt{d})}}^{(\mathtt{n,d})}$, and decoy tumor HLA sequences $S_1^{(\mathtt{t,d})}, \ldots, S_{\nu^{(\mathtt{d})}}^{(\mathtt{t,d})}$, where $\nu^{(\mathtt{d})}$ is a hyperparameter of the number of the decoy parameters. Here, the sequences of $R_1^{(\mathrm{r})}$ and $R_2^{(\mathrm{r})}$ are the MSAs of the HLA types. $S_1^{(\mathrm{n,r})}$ and $S_2^{(\mathrm{n,r})}$ are used to consider germline variants in $R_1^{(\mathrm{r})}$ and $R_2^{(\mathrm{r})}$, and $S_1^{(\mathtt{t,r})}$ and $S_2^{(\mathtt{t,r})}$ are used to reflect somatic mutations. As in the case of ALPHLARD, these decoy parameters are essential to make a robust inference because their presence can reduce the influence of misclassified reads that were actually produced by other HLA genes and HLA 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 $(R_1^{(\mathrm{r})}, R_2^{(\mathrm{r})}, R_1^{(\mathtt{d})}, \ldots, R_{\nu^{(\mathtt{d})}}^{(\mathtt{d})})$, $(S_1^{(\mathrm{n,r})}, S_2^{(\mathrm{n,r})}, S_1^{(\mathtt{n,d})}, \ldots, S_{\nu^{(\mathtt{d})}}^{(\mathtt{n,d})})$, and $(S_1^{(\mathtt{t,r})}, S_2^{(\mathtt{t,r})}, S_1^{(\mathtt{t,d})}, \ldots, S_{\nu^{(\mathtt{d})}}^{(\mathtt{t,d})})$, respectively. In addition, in some cases, $(S_1, ..., S_{2\nu^{(\mathtt{d})}+4})$ is used instead of $(S_1^{(\mathtt{n})}, \ldots, S_{\nu^{(\mathtt{d})}+2}^{(\mathtt{n})}, S_1^{(\mathtt{t})}, \ldots, S_{\nu^{(\mathtt{d})}+2}^{(\mathtt{t})})$. We also denote the n^{th} base or gap of


$\mathcal{R}^{(\mathtt{d})}$
$R_2^{(\mathtt{r})}$
$R_1^{(\mathtt{r})}$
$\mathcal{S}^{(\mathtt{n,d})}$
$S_2^{(\mathtt{n,r})}$
$S_1^{(\mathtt{n,r})}$
$\mathcal{X}^{(\mathtt{n})}$
$\mathcal{I}^{(\mathtt{n})}$
$\mathcal{V}^{(\mathtt{d})}$
$V_2^{(\mathtt{r})}$
$V_1^{(\mathtt{r})}$
$\mathcal{S}^{(\mathtt{t,d})}$
$S_2^{(\mathtt{t,r})}$
$S_1^{(\mathtt{t,r})}$
$\mathcal{X}^{(\mathtt{t})}$
$\mathcal{I}^{(\mathtt{t})}$
$\mathcal{F}^{(\mathtt{d})}$
$F_2^{(\mathtt{r})}$
G
$F_1^{(\mathtt{r})}$

$R_1^{(\mathtt{r})}$, $R_2^{(\mathtt{r})}$: HLA types
$\mathcal{R}^{(\mathtt{d})}$: decoy HLA types
$S_1^{(\mathtt{n,r})}$, $S_2^{(\mathtt{n,r})}$: normal HLA sequences
$\mathcal{S}^{(\mathtt{n,d})}$: decoy normal HLA sequences
$S_1^{(\mathtt{t,r})}$, $S_2^{(\mathtt{t,r})}$: tumor HLA sequences
$\mathcal{S}^{(\mathtt{t,d})}$: decoy tumor HLA sequences
$\mathcal{X}^{(\mathtt{n})}$: normal realigned read pairs
$\mathcal{X}^{(\mathtt{t})}$: tumor realigned read pairs
$\mathcal{I}^{(\mathtt{n})}$: variables that indicate which HLA sequence produced each normal read pair
$\mathcal{I}^{(\mathtt{t})}$: variables that indicate which HLA sequence produced each tumor read pair
V_1, V_2, $\mathcal{V}^{(\mathtt{d})}$: variables that indicate whether each HLA sequence is valid
F_1, F_2, $\mathcal{F}^{(\mathtt{d})}$: 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

Figure 5.1: Graphical representation of ALPHLARD-NT.

R_m in the combined MSA by $R_{m,n}$, and the n^{th} base or gap of S_m by $S_{m,n}$. Note that all of the R_m's and S_m's have the same sequence length as aligned sequences in the combined MSA, which we denote by N. 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, $I_i^{(\mathtt{n})} = m$ means that $x_i^{(\mathtt{n})}$ was produced by S_m, and $I_i^{(\mathtt{t})} = m$ means that $x_i^{(\mathtt{t})}$ was produced by S_m. Similarly to ALPHLARD, the influence of misclassified reads from other HLA genes and HLA pseudogenes can be ignored by assigning the corresponding indicator variables to decoy HLA sequences. 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.

In addition to parameters for the tumor sample, we also introduce parameters of the prior distribution of indicator variables $I_i^{(\mathtt{n})}$'s and $I_i^{(\mathtt{t})}$'s. Each of the indicator variables is assumed to be independently generated from a distribution that is 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 $(V_1^{(\mathtt{r})}, V_2^{(\mathtt{r})}, V_1^{(\mathtt{d})}, \ldots, V_{\nu^{(\mathtt{d})}}^{(\mathtt{d})})$. Here, F_m is a positive real parameter that expresses the likelihood that a read is produced by $S_m^{(\mathtt{n})}$ or $S_m^{(\mathtt{t})}$. 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 $V_{m,n}$ is a validity flag for $S_{m,n}^{(\mathtt{n})}$ and $S_{m,n}^{(\mathtt{t})}$; in other words, $V_{m,n}$ takes 0 or 1, and indicates whether $S_{m,n}^{(\mathtt{n})}$ and $S_{m,n}^{(\mathtt{t})}$ are valid, as described in more detail below.

Then, the posterior probability of the parameters is given by

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

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

$p(\mathcal{X}^{(\mathtt{n})} \mid \mathcal{S}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{n})})$ and $p(\mathcal{X}^{(\mathtt{t})} \mid \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{t})})$ are the likelihoods of normal and tumor realigned read pairs and are defined by

$$
\begin{aligned}
p(\mathcal{X}^{(\mathtt{n})} \mid \mathcal{S}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{n})}) &= \prod_i \prod_j \prod_n p(x_{i,j,n}^{(\mathtt{n})} \mid S_{I_i^{(\mathtt{n})},n}), \\
p(\mathcal{X}^{(\mathtt{t})} \mid \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{t})}) &= \prod_i \prod_j \prod_n p(x_{i,j,n}^{(\mathtt{t})} \mid S_{I_i^{(\mathtt{t})},n}),
\end{aligned}
$$

where

$$
\begin{aligned}
&p(x_{i,j,n} \mid S_{m,n} \in B) \\
&\quad = \begin{cases}
(1 - \pi^{(\mathtt{e,d})})(1 - \pi^{(\mathtt{e,N})})(1 - p_{i,j,n}) & (\text{if } x_{i,j,n} = S_{m,n}) \\
(1 - \pi^{(\mathtt{e,d})})(1 - \pi^{(\mathtt{e,N})})\frac{p_{i,j,n}}{3} & (\text{if } x_{i,j,n} \in B \text{ and } x_{i,j,n} \neq S_{m,n}) \\
(1 - \pi^{(\mathtt{e,d})})\pi^{(\mathtt{e,N})} & (\text{if } x_{i,j,n} = \mathtt{N}) \\
\pi^{(\mathtt{e,d})} & (\text{if } x_{i,j,n} = \mathtt{-})
\end{cases}, \\
&p(x_{i,j,n} \mid S_{m,n} = \mathtt{-}) \\
&\quad = \begin{cases}
\pi^{(\mathtt{e,i})}(1 - \pi^{(\mathtt{e,N})})\frac{1}{4} & (\text{if } x_{i,j,n} \in B) \\
\pi^{(\mathtt{e,i})}\pi^{(\mathtt{e,N})} & (\text{if } x_{i,j,n} = \mathtt{N}) \\
1 - \pi^{(\mathtt{e,i})} & (\text{if } x_{i,j,n} = \mathtt{-})
\end{cases}, \\
&p(x_{i,j,n} \mid S_{m,n} = \mathtt{N}) \\
&\quad = \begin{cases}
(1 - \pi^{(\mathtt{e,N})})\frac{1}{5} & (\text{if } x_{i,j,n} \in B \text{ or } x_{i,j,n}\mathtt{-}) \\
\pi^{(\mathtt{e,N})} & (\text{if } x_{i,j,n} = \mathtt{N})
\end{cases}.
\end{aligned}
$$

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

$p(\mathcal{S}^{(\mathtt{t})} \mid \mathcal{S}^{(\mathtt{n})})$ is the prior probability of tumor HLA sequences and is defined by

$$
p(\mathcal{S}^{(\mathtt{t})} \mid \mathcal{S}^{(\mathtt{n})}) = \prod_m \prod_n p(S_{m,n}^{(\mathtt{t})} \mid S_{m,n}^{(\mathtt{n})}),
$$

where

$$
\begin{aligned}
&p(S^{(\mathtt{t})}_{m,n} \mid S^{(\mathtt{n})}_{m,n} \in B) \\
&\quad = \begin{cases}
(1-\pi^{(\mathtt{s},\mathtt{N})})(1-\pi^{(\mathtt{s},\mathtt{d})})(1-\pi^{(\mathtt{s},\mathtt{s})}) & (\text{if } S^{(\mathtt{t})}_{m,n} = S^{(\mathtt{n})}_{m,n}) \\
(1-\pi^{(\mathtt{s},\mathtt{N})})(1-\pi^{(\mathtt{s},\mathtt{d})})\frac{\pi^{(\mathtt{s},\mathtt{s})}}{3} & (\text{if } S^{(\mathtt{t})}_{m,n} \in B \text{ and } S^{(\mathtt{t})}_{m,n} \neq S^{(\mathtt{n})}_{m,n}) \\
(1-\pi^{(\mathtt{s},\mathtt{N})})\pi^{(\mathtt{s},\mathtt{d})} & (\text{if } S^{(\mathtt{t})}_{m,n} = \mathtt{-}) \\
\pi^{(\mathtt{s},\mathtt{N})} & (\text{if } S^{(\mathtt{t})}_{m,n} = \mathtt{N})
\end{cases}, \\
&p(S^{(\mathtt{t})}_{m,n} \mid S^{(\mathtt{n})}_{m,n} = \mathtt{-}) \\
&\quad = \begin{cases}
(1-\pi^{(\mathtt{s},\mathtt{N})})\pi^{(\mathtt{s},\mathtt{i})}\frac{1}{4} & (\text{if } S^{(\mathtt{t})}_{m,n} \in B) \\
(1-\pi^{(\mathtt{s},\mathtt{N})})(1-\pi^{(\mathtt{s},\mathtt{i})}) & (\text{if } S^{(\mathtt{t})}_{m,n} = \mathtt{-}) \\
\pi^{(\mathtt{s},\mathtt{N})} & (\text{if } S^{(\mathtt{t})}_{m,n} = \mathtt{N})
\end{cases}, \\
&p(S^{(\mathtt{t})}_{m,n} \mid S^{(\mathtt{n})}_{m,n} = \mathtt{N}) \\
&\quad = \begin{cases}
(1-\pi^{(\mathtt{s},\mathtt{N})})\frac{1}{5} & (\text{if } S^{(\mathtt{t})}_{m,n} \in B \text{ or } S^{(\mathtt{t})}_{m,n} = \mathtt{-}) \\
\pi^{(\mathtt{s},\mathtt{N})} & (\text{if } S^{(\mathtt{t})}_{m,n} = \mathtt{N})
\end{cases}.
\end{aligned}
$$

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

$p(\mathcal{S}^{(\mathtt{n})} \mid \mathcal{R})$ is the prior probability of normal HLA sequences and is defined by

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

where

$$
\begin{aligned}
&p(S^{(\mathtt{n},\mathtt{r})}_{m,n} \mid R^{(\mathtt{r})}_{m,n} \in B, R^{(\mathtt{r})}_{m,n} \text{ is original}) \\
&\quad = \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} & \left(\begin{array}{l} \text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} \in B \\ \quad \text{and } S^{(\mathtt{n},\mathtt{r})}_{m,n} \neq R^{(\mathtt{r})}_{m,n} \end{array}\right) \\
(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} = \mathtt{-}) \\
\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} \mid R^{(\mathtt{r})}_{m,n} = \mathtt{-}, R^{(\mathtt{r})}_{m,n} \text{ is original}) \\
&\quad = \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} = \mathtt{-}) \\
\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} \mid R^{(\mathtt{r})}_{m,n} = \mathtt{N}, R^{(\mathtt{r})}_{m,n} \text{ is original}) \\
&\quad = \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} = \mathtt{-}) \\
\pi^{(\mathtt{g},\mathtt{r},\mathtt{o},\mathtt{N})} & (\text{if } S^{(\mathtt{n},\mathtt{r})}_{m,n} = \mathtt{N})
\end{cases},
\end{aligned}
$$

$$p(S^{(\mathtt{n},\mathtt{r})}_{m,n} \mid R^{(\mathtt{r})}_{m,n} \in B, R^{(\mathtt{r})}_{m,n} \text{ is imputed})$$

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

$$p(S^{(\mathtt{n},\mathtt{r})}_{m,n} \mid R^{(\mathtt{r})}_{m,n} = \mathtt{-}, R^{(\mathtt{r})}_{m,n} \text{ is imputed})$$

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

$$p(S^{(\mathtt{n},\mathtt{r})}_{m,n} \mid R^{(\mathtt{r})}_{m,n} = \mathtt{N}, R^{(\mathtt{r})}_{m,n} \text{ is imputed})$$

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

$$p(S^{(\mathtt{n},\mathtt{d})}_{m,n} \mid R^{(\mathtt{d})}_{m,n} \in B, R^{(\mathtt{d})}_{m,n} \text{ is original})$$

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

$$p(S^{(\mathtt{n},\mathtt{d})}_{m,n} \mid R^{(\mathtt{d})}_{m,n} = \mathtt{-}, R^{(\mathtt{d})}_{m,n} \text{ is original})$$

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

$$p(S^{(\mathtt{n},\mathtt{d})}_{m,n} \mid R^{(\mathtt{d})}_{m,n} = \mathtt{N}, R^{(\mathtt{d})}_{m,n} \text{ is original})$$

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

$$\begin{aligned}
&p(S^{(\mathtt{n,d})}_{m,n} \mid R^{(\mathtt{d})}_{m,n} \in B, R^{(\mathtt{d})}_{m,n} \text{ is imputed}) \\
&\quad = \begin{cases}
(1-\pi^{(\mathtt{g,d,i,N})})(1-\pi^{(\mathtt{g,d,i,d})})(1-\pi^{(\mathtt{g,d,i,s})}) & (\text{if } S^{(\mathtt{n,d})}_{m,n} = R^{(\mathtt{d})}_{m,n}) \\
(1-\pi^{(\mathtt{g,d,i,N})})(1-\pi^{(\mathtt{g,d,i,d})})\frac{\pi^{(\mathtt{g,d,i,s})}}{3} & \begin{pmatrix} \text{if } S^{(\mathtt{n,d})}_{m,n} \in B \\ \text{and } S^{(\mathtt{n,d})}_{m,n} \neq R^{(\mathtt{d})}_{m,n} \end{pmatrix} \\
(1-\pi^{(\mathtt{g,d,i,N})})\pi^{(\mathtt{g,d,i,d})} & (\text{if } S^{(\mathtt{n,d})}_{m,n} = \texttt{-}) \\
\pi^{(\mathtt{g,d,i,N})} & (\text{if } S^{(\mathtt{n,d})}_{m,n} = \texttt{N})
\end{cases}, \\
&p(S^{(\mathtt{n,d})}_{m,n} \mid R^{(\mathtt{d})}_{m,n} = \texttt{-}, R^{(\mathtt{d})}_{m,n} \text{ is imputed}) \\
&\quad = \begin{cases}
(1-\pi^{(\mathtt{g,d,i,N})})\pi^{(\mathtt{g,d,i,i})}\frac{1}{4} & (\text{if } S^{(\mathtt{n,d})}_{m,n} \in B) \\
(1-\pi^{(\mathtt{g,d,i,N})})(1-\pi^{(\mathtt{g,d,i,i})}) & (\text{if } S^{(\mathtt{n,d})}_{m,n} = \texttt{-}) \\
\pi^{(\mathtt{g,d,i,N})} & (\text{if } S^{(\mathtt{n,d})}_{m,n} = \texttt{N})
\end{cases}, \\
&p(S^{(\mathtt{n,d})}_{m,n} \mid R^{(\mathtt{d})}_{m,n} = \texttt{N}, R^{(\mathtt{d})}_{m,n} \text{ is imputed}) \\
&\quad = \begin{cases}
(1-\pi^{(\mathtt{g,d,i,N})})\frac{1}{5} & (\text{if } S^{(\mathtt{n,d})}_{m,n} \in B \text{ or } S^{(\mathtt{n,d})}_{m,n} = \texttt{-}) \\
\pi^{(\mathtt{g,d,i,N})} & (\text{if } S^{(\mathtt{n,d})}_{m,n} = \texttt{N})
\end{cases}.
\end{aligned}$$

Here, $\pi^{(\mathtt{g,r,o,s})}$, $\pi^{(\mathtt{g,r,o,d})}$, $\pi^{(\mathtt{g,r,o,i})}$, $\pi^{(\mathtt{g,r,o,N})}$ are hyperparameters of the probabilities of a germline substitution, a germline deletion, a germline insertion, and an N, respectively, in a non-decoy normal HLA sequence at the position where the reference is an original base. Similarly, $\pi^{(\mathtt{g,r,i,s})}$, $\pi^{(\mathtt{g,r,i,d})}$, $\pi^{(\mathtt{g,r,i,i})}$, $\pi^{(\mathtt{g,r,i,N})}$ are hyperparameters of the probabilities of a germline substitution, a germline deletion, a germline insertion, and an N, respectively, in a non-decoy normal HLA sequence at the position where the reference is an imputed base. Also, $\pi^{(\mathtt{g,d,o,s})}$, $\pi^{(\mathtt{g,d,o,d})}$, $\pi^{(\mathtt{g,d,o,i})}$, $\pi^{(\mathtt{g,d,o,N})}$ are hyperparameters of the probabilities of a germline substitution, a germline deletion, a germline insertion, and an N, respectively, in a decoy normal HLA sequence at the position where the reference is an original base. Lastly, $\pi^{(\mathtt{g,d,i,s})}$, $\pi^{(\mathtt{g,d,i,d})}$, $\pi^{(\mathtt{g,d,i,i})}$, $\pi^{(\mathtt{g,d,i,N})}$ are hyperparameters of the probabilities of a germline substitution, a germline deletion, a germline insertion, and an N, respectively, in a decoy normal HLA sequence at the position where the reference is an imputed base. These hyperparameters determine how likely germline mutations in HLA genes are to occur. The probabilities for an imputed reference base should be larger than or equal to 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 or equal to those for a non-decoy normal HLA sequence to achieve robustness against misclassified reads. Moreover, $S^{(\mathtt{n})}_{m,n}$ can be N, and tends to be N when $S_{m,n}$ cannot be uniquely determined, which can occur when the realigned read pairs $\mathcal{X}$ includes a number of misclassified reads from other HLA genes and pseudogenes that cannot be ignored. In this case, allowing $S^{(\mathtt{n})}_{m,n}$ to be N can make HLA genotyping robuster.

$p(\mathcal{R})$ is the prior probability of HLA types and is defined by

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

Here, $p(R^{(\mathtt{r})}_m)$ is the prior probability of the HLA type and is given from the result of Section 3.2.1. On the other hand, $p(R^{(\mathtt{d})}_m)$ is the prior probability of the decoy HLA type, which we assume as constant.

$p(\mathcal{I}^{(\mathtt{n})} \mid \mathcal{F}, \mathcal{V})$ is the prior probability of normal indicator variables and is defined by

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

where

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

This formula means that the read cannot be produced by the HLA sequence without a valid position covered by the read, which is controlled by $\mathcal{V}$.

Similarly, $p(\mathcal{I}^{(\mathtt{t})} \mid \mathcal{F}, G, \mathcal{V})$ is the prior probability of tumor indicator variables and is defined by

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

where

$$\begin{aligned} p(I_i^{(\mathtt{t})} = m \in M^{(\mathtt{n})} \mid \mathcal{F}, G, \mathcal{V}) &\propto \left(\max_{n \in \bigcup_j r_{i,j}^{(\mathtt{t})}} V_m \right) F_m G, \\ p(I_i^{(\mathtt{t})} = m \in M^{(\mathtt{t})} \mid \mathcal{F}, G, \mathcal{V}) &\propto \left(\max_{n \in \bigcup_j r_{i,j}^{(\mathtt{t})}} V_{m-(\nu^{(\mathtt{d})}+2)} \right) F_{m-(\nu^{(\mathtt{d})}+2)}, \\ M^{(\mathtt{n})} &= \{1, \ldots, \nu^{(\mathtt{d})} + 2\}, \\ M^{(\mathtt{t})} &= \{\nu^{(\mathtt{d})} + 3, \ldots, 2\nu^{(\mathtt{d})} + 4\}. \end{aligned}$$

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. Further, matched normal-tumor HLA sequences $S_m^{(\mathtt{n})}$ and $S_m^{(\mathtt{t})}$ share V_m and F_m.

p($\mathcal{F}$) is the prior probability of likelihoods of indicator variables and is defined by

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

where

$$\begin{aligned} p(F_m^{(\mathtt{r})}) &= \mathcal{LN}(F_m^{(\mathtt{r})} \mid \mu^{(\mathtt{f,r})}, (\sigma^{(\mathtt{f,r})})^2), \\ p(F_m^{(\mathtt{d})}) &= \mathcal{LN}(F_m^{(\mathtt{d})} \mid \mu^{(\mathtt{f,d})}, (\sigma^{(\mathtt{f,d})})^2). \end{aligned}$$

Here, $\mathcal{LN}$ is a log-normal distribution, $\mu^{(\mathtt{f,r})}$ and $(\sigma^{(\mathtt{f,r})})^2$ are hyperparameters of the mean and the variance for non-decoy parameters, and $\mu^{(\mathtt{f,d})}$ and $(\sigma^{(\mathtt{f,d})})^2$ are hyperparameters of the mean and the variance for decoy parameters. $\mu^{(\mathtt{f,d})}$ should be smaller than or equal to $\mu^{(\mathtt{f,r})}$ because realigned read pairs that are mapped to decoy HLA sequences should be basically removed at the extraction and classification step described in Chapter 3.

$p(G)$ is the prior probability of the ratio of normal cells contained in the tumor sample and is defined by

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

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

$p(\mathcal{V})$ is the prior probability of validity flags and is defined by

$$\begin{aligned} p(\mathcal{V}) &= \left(\prod_m p(V^{(\mathtt{r})})\right)\left(\prod_m p(V^{(\mathtt{d})})\right) \\ &= \left(\prod_m \prod_n p(V^{(\mathtt{r})}_{m,n})\right)\left(\prod_m \prod_n p(V^{(\mathtt{d})}_{m,n} \mid V^{(\mathtt{d})}_{m,n-1})\right), \end{aligned}$$

where

$$\begin{aligned} p(V^{(\mathtt{r})}_{m,n}) &= \begin{cases} 0 & (\text{if } V^{(\mathtt{r})}_{m,n} = 0) \\ 1 & (\text{if } V^{(\mathtt{r})}_{m,n} = 1) \end{cases}, \\ p(V^{(\mathtt{d})}_{m,n} \mid V^{(\mathtt{d})}_{m,n-1} = 0) &= \begin{cases} 1 - \pi^{(\mathtt{v,o})} & (\text{if } V^{(\mathtt{d})}_{m,n} = 0) \\ \pi^{(\mathtt{v,o})} & (\text{if } V^{(\mathtt{d})}_{m,n} = 1) \end{cases}, \\ p(V^{(\mathtt{d})}_{m,n} \mid V^{(\mathtt{d})}_{m,n-1} = 1) &= \begin{cases} 1 - \pi^{(\mathtt{v,e})} & (\text{if } V^{(\mathtt{d})}_{m,n} = 0) \\ \pi^{(\mathtt{v,e})} & (\text{if } V^{(\mathtt{d})}_{m,n} = 1) \end{cases}. \end{aligned}$$

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

5.4 MCMC Sampling of Parameters

Similarly to ALPHLARD, we use MCMC to sample parameters from the posterior distribution with parallel tempering. Gibbs sampling is primarily used to sample all parameters except for F_m's, G, and V_m's. F_m's, G, and V_m's are sampled using the Metropolis-Hastings algorithm. In addition, we also periodically use the Metropolis-Hastings algorithm that allows parameters to move from mode to mode. In the following sections, we introduce how parameter sampling is conducted.

5.4.1 Gibbs Sampling of HLA Types $\mathcal{R}$

Gibbs Sampling of HLA Types $\mathcal{R}$ is similar to that in ALPHLARD. The conditional distribution of HLA types is given by

$$\begin{aligned} & p(\mathcal{R} \mid \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}, G, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})}) \\ & \quad \propto p(\mathcal{S}^{(\mathtt{n})} \mid \mathcal{R}) p(\mathcal{R}) \\ & \quad = \prod_m \left(\prod_n p(S^{(\mathtt{n})}_{m,n} \mid R_{m,n})\right) p(R_m). \end{aligned}$$

Therefore, each HLA type R_m can be sampled independently from the probability distribution in proportion to

$$p(S^{(\mathtt{n})}_m \mid R_m) p(R_m),$$

using Gibbs sampling. The Gibbs sampling of each HLA type is shown in Algorithm 5.1.

Algorithm 5.1 Gibbs sampling of each HLA type in ALPHLARD-NT

```
Input:
    $S_m^{(\mathtt{n})}$: the $m^{\text{th}}$ normal HLA sequence
Output:
    $R_m$: the $m^{\text{th}}$ HLA type

$T \leftarrow$ a set of HLA types
for all $t \in T$ do
    $p_t \leftarrow 1$
end for
for all $t \in T$ do
    for $n \leftarrow 1$ to $N$ do
        $p_t \leftarrow p_t \times p(S_{m,n}^{(\mathtt{n})} \mid R_{m,n}, R_m = t)$
    end for
    $p_t \leftarrow p_t \times p(R_m = t)$
end for
Sample $R_m$ with probability in proportion to $\boldsymbol{p}$
return $R_m$
```

5.4.2 Gibbs Sampling of normal HLA Sequences $\mathcal{S}^{(\mathtt{n})}$

The conditional distribution of normal HLA sequences is given by

$$\begin{aligned}
&p(\mathcal{S}^{(\mathtt{n})} \mid \mathcal{R}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}, G, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})}) \\
&\qquad \propto p(\mathcal{X}^{(\mathtt{n})} \mid \mathcal{S}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{n})}) p(\mathcal{X}^{(\mathtt{t})} \mid \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{t})}) p(\mathcal{S}^{(\mathtt{t})} \mid \mathcal{S}^{(\mathtt{n})}) p(\mathcal{S}^{(\mathtt{n})} \mid \mathcal{R}) \\
&\qquad = \left(\prod_i \prod_j \prod_n p(x_{i,j,n}^{(\mathtt{n})} \mid S_{I_i^{(\mathtt{n})},n}) \right) \left(\prod_i \prod_j \prod_n p(x_{i,j,n}^{(\mathtt{t})} \mid S_{I_i^{(\mathtt{t})},n}) \right) \\
&\qquad\qquad \times \left(\prod_m \prod_n p(S_{m,n}^{(\mathtt{t})} \mid S_{m,n}^{(\mathtt{n})}) \right) \left(\prod_m \prod_n p(S_{m,n}^{(\mathtt{n})} \mid R_{m,n}) \right).
\end{aligned}$$

Here, the likelihoods of normal realigned read pairs and tumor realigned read pairs can be rewritten as

$$\begin{aligned}
\prod_i \prod_j \prod_n p(x_{i,j,n}^{(\mathtt{n})} \mid S_{I_i^{(\mathtt{n})},n}) &= \prod_m \prod_n \prod_{i:I_i^{(\mathtt{n})}=m} \prod_j p(x_{i,j,n}^{(\mathtt{n})} \mid S_{m,n}^{(\mathtt{n})}), \\
\prod_i \prod_j \prod_n p(x_{i,j,n}^{(\mathtt{t})} \mid S_{I_i^{(\mathtt{t})},n}) &\propto \prod_m \prod_n \prod_{i:I_i^{(\mathtt{t})}=m} \prod_j p(x_{i,j,n}^{(\mathtt{t})} \mid S_{m,n}^{(\mathtt{n})}).
\end{aligned}$$

As a result, The conditional distribution of normal HLA sequences can be calculated by

$$\begin{aligned}
&p(\mathcal{S}^{(\mathtt{n})} \mid \mathcal{R}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})}) \\
&\qquad \propto \prod_m \prod_n \left(\prod_{i:I_i^{(\mathtt{n})}=m} \prod_j p(x_{i,j,n}^{(\mathtt{n})} \mid S_{m,n}^{(\mathtt{n})}) \right) \left(\prod_{i:I_i^{(\mathtt{t})}=m} \prod_j p(x_{i,j,n}^{(\mathtt{t})} \mid S_{m,n}^{(\mathtt{n})}) \right) \\
&\qquad\qquad \times p(S_{m,n}^{(\mathtt{t})} \mid S_{m,n}^{(\mathtt{n})}) p(S_{m,n}^{(\mathtt{n})} \mid R_{m,n}).
\end{aligned}$$

Therefore, each normal HLA base $S^{(\mathtt{n})}_{m,n}$ can be sampled independently from the probability distribution in proportion to

$$\left(\prod_{i:I^{(\mathtt{n})}_i=m}\prod_j p(x^{(\mathtt{n})}_{i,j,n} \mid S^{(\mathtt{n})}_{m,n})\right)\left(\prod_{i:I^{(\mathtt{t})}_i=m}\prod_j p(x^{(\mathtt{t})}_{i,j,n} \mid S^{(\mathtt{n})}_{m,n})\right)$$
$$\times\, p(S^{(\mathtt{t})}_{m,n} \mid S^{(\mathtt{n})}_{m,n})p(S^{(\mathtt{n})}_{m,n} \mid R_{m,n}),$$

using Gibbs sampling. The Gibbs sampling of each normal HLA base is shown in Algorithm 5.2.

5.4.3 Gibbs Sampling of tumor HLA Sequences $\mathcal{S}^{(\mathtt{t})}$

The conditional distribution of tumor HLA sequences is given by

$$\begin{aligned}
&p(\mathcal{S}^{(\mathtt{t})} \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{F}, G, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})\\
&\qquad \propto p(\mathcal{X}^{(\mathtt{t})} \mid \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{t})})p(\mathcal{S}^{(\mathtt{t})} \mid \mathcal{S}^{(\mathtt{n})})\\
&\qquad = \left(\prod_i\prod_j\prod_n p(x^{(\mathtt{t})}_{i,j,n} \mid S_{I^{(\mathtt{t})}_i,n})\right)\left(\prod_m\prod_n p(S^{(\mathtt{t})}_{m,n} \mid S^{(\mathtt{n})}_{m,n})\right).
\end{aligned}$$

Here, the likelihood of tumor realigned read pairs can be rewritten as

$$\prod_i\prod_j\prod_n p(x^{(\mathtt{t})}_{i,j,n} \mid S_{I^{(\mathtt{t})}_i,n}) \propto \prod_m\prod_n \prod_{i:I^{(\mathtt{t})}_i=m+\nu^{(\mathtt{d})}+2}\prod_j p(x^{(\mathtt{t})}_{i,j,n} \mid S^{(\mathtt{t})}_{m,n}),$$

where the addition by $\nu^{(\mathtt{d})}+2$ means that the reads are produced by not normal HLA sequences but tumor HLA sequences. As a result, The conditional distribution of tumor HLA sequences can be calculated by

$$\begin{aligned}
&p(\mathcal{S}^{(\mathtt{n})} \mid \mathcal{R}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})\\
&\qquad \propto \prod_m\prod_n\left(\prod_{i:I^{(\mathtt{t})}_i=m+\nu^{(\mathtt{d})}+2}\prod_j p(x^{(\mathtt{t})}_{i,j,n} \mid S^{(\mathtt{t})}_{m,n})\right)p(S^{(\mathtt{t})}_{m,n} \mid S^{(\mathtt{n})}_{m,n}).
\end{aligned}$$

Therefore, each tumor HLA base $S^{(\mathtt{t})}_{m,n}$ can be sampled independently from the probability distribution in proportion to

$$\left(\prod_{i:I^{(\mathtt{t})}_i=m+\nu^{(\mathtt{d})}+2}\prod_j p(x^{(\mathtt{t})}_{i,j,n} \mid S^{(\mathtt{t})}_{m,n})\right)p(S^{(\mathtt{t})}_{m,n} \mid S^{(\mathtt{n})}_{m,n}),$$

using Gibbs sampling. The Gibbs sampling of each tumor HLA base is shown in Algorithm 5.3.

Algorithm 5.2 Gibbs sampling of each normal HLA base in ALPHLARD-NT

```
Input:
    $\mathcal{R}$: HLA types
    $\mathcal{I}^{(\mathtt{n})}$: normal indicator variables
    $\mathcal{I}^{(\mathtt{t})}$: tumor indicator variables
    $\mathcal{X}^{(\mathtt{n})}$: normal realigned read pairs
    $\mathcal{X}^{(\mathtt{t})}$: tumor realigned read pairs
Output:
    $\mathcal{S}^{(\mathtt{n})}$: normal HLA sequences

$K^{(\mathtt{n})} \leftarrow$ the number of normal realigned read pairs
$K^{(\mathtt{t})} \leftarrow$ the number of tumor realigned read pairs
$B \leftarrow \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}, \mathtt{-}, \mathtt{N}\}$
for $m \leftarrow 1$ to $\nu^{(\mathtt{d})} + 2$ do
    for $n \leftarrow 1$ to $N$ do
        for $b \in B$ do
            $p_{m,n,b} \leftarrow 1$
        end for
    end for
end for
for $s \in \{\mathtt{n}, \mathtt{t}\}$ do
    for $i \leftarrow 1$ to $K^{(s)}$ do
        continue if $I_i^{(s)} > \nu^{(\mathtt{d})} + 2$
        $m \leftarrow I_i^{(s)}$
        $J \leftarrow \begin{cases} 1 & (\text{if } x_i^{(s)} \text{ is a paired read}) \\ 2 & (\text{if } x_i^{(s)} \text{ is an unpaired read}) \end{cases}$
        for $j \leftarrow 1$ to $J$ do
            $r \leftarrow$ a set of positions covered by the read $x_{i,j}^{(s)}$
            for $n \in r$ do
                for $b \in B$ do
                    $p_{m,n,b} \leftarrow p_{m,n,b} \times p(x_{i,j,n}^{(s)} \mid S_{m,n}^{(\mathtt{n})} = b)$
                end for
            end for
        end for
    end for
end for
for $m \leftarrow 1$ to $\nu^{(\mathtt{d})} + 2$ do
    for $n \leftarrow 1$ to $N$ do
        for $b \in B$ do
            $p_{m,n,b} \leftarrow p_{m,n,b} \times p(S_{m,n}^{(\mathtt{t})} \mid S_{m,n}^{(\mathtt{n})} = b)$
            $p_{m,n,b} \leftarrow p_{m,n,b} \times p(S_{m,n}^{(\mathtt{n})} = b \mid R_{m,n})$
        end for
    end for
end for
for $m \leftarrow 1$ to $\nu^{(\mathtt{d})} + 2$ do
    for $n \leftarrow 1$ to $N$ do
        Sample $S_{m,n}^{(\mathtt{n})}$ with probability in proportion to $\boldsymbol{p_{m,n}}$
    end for
end for
return $\mathcal{S}^{(\mathtt{n})}$
```

Algorithm 5.3 Gibbs sampling of each tumor HLA base in ALPHLARD-NT

```
Input:
    $\mathcal{S}^{(\mathtt{n})}$: normal HLA sequences
    $\mathcal{I}^{(\mathtt{t})}$: tumor indicator variables
    $\mathcal{X}^{(\mathtt{t})}$: tumor realigned read pairs
Output:
    $\mathcal{S}^{(\mathtt{t})}$: tumor HLA sequences

$K^{(\mathtt{t})} \leftarrow$ the number of normal realigned read pairs
$B \leftarrow \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}, \mathtt{-}, \mathtt{N}\}$
for $m \leftarrow 1$ to $\nu^{(\mathtt{d})} + 2$ do
    for $n \leftarrow 1$ to $N$ do
        for $b \in B$ do
            $p_{m,n,b} \leftarrow 1$
        end for
    end for
end for
for $i \leftarrow 1$ to $K$ do
    continue if $I_i^{(\mathtt{t})} \leq \nu^{(\mathtt{d})} + 2$
    $m \leftarrow I_i^{(\mathtt{t})} - (\nu^{(\mathtt{d})} + 2)$
    $J \leftarrow \begin{cases} 1 & (\text{if } x_i^{(\mathtt{t})} \text{ is a paired read}) \\ 2 & (\text{if } x_i^{(\mathtt{t})} \text{ is an unpaired read}) \end{cases}$
    for $j \leftarrow 1$ to $J$ do
        $r \leftarrow$ a set of positions covered by the read $x_{i,j}^{(\mathtt{t})}$
        for $n \in r$ do
            for $b \in B$ do
                $p_{m,n,b} \leftarrow p_{m,n,b} \times p(x_{i,j,n}^{(\mathtt{t})} \mid S_{m,n}^{(\mathtt{t})} = b)$
            end for
        end for
    end for
end for
for $m \leftarrow 1$ to $\nu^{(\mathtt{d})} + 2$ do
    for $n \leftarrow 1$ to $N$ do
        for $b \in B$ do
            $p_{m,n,b} \leftarrow p_{m,n,b} \times p(S_{m,n}^{(\mathtt{t})} = b \mid S_{m,n}^{(\mathtt{n})})$
        end for
    end for
end for
for $m \leftarrow 1$ to $\nu^{(\mathtt{d})} + 2$ do
    for $n \leftarrow 1$ to $N$ do
        Sample $S_{m,n}^{(\mathtt{t})}$ with probability in proportion to $\boldsymbol{p_{m,n}}$
    end for
end for
return $\mathcal{S}^{(\mathtt{t})}$
```

Algorithm 5.4 Gibbs sampling of each normal indicator variable in ALPHLARD-NT

```
Input:
    $\mathcal{S}^{(\mathtt{n})}$: normal HLA sequences
    $\mathcal{F}$: likelihoods of indicator variables
    $\mathcal{V}$: validity flags
    $x_i^{(\mathtt{n})}$: the $i^{\mathrm{th}}$ normal realigned read pair
Output:
    $I_i^{(\mathtt{n})}$: the $i^{\mathrm{th}}$ normal indicator variable

$J \leftarrow \begin{cases} 1 & (\text{if } x_i^{(\mathtt{n})} \text{ is a paired read}) \\ 2 & (\text{if } x_i^{(\mathtt{n})} \text{ is an unpaired read}) \end{cases}$
for $m \leftarrow 1$ to $\nu^{(\mathtt{d})} + 2$ do
    $p_m \leftarrow 1$
end for
for $j \leftarrow 1$ to $J$ do
    $r \leftarrow$ a set of positions covered by the read $x_{i,j}^{(\mathtt{n})}$
    for $m \leftarrow 1$ to $\nu^{(\mathtt{d})} + 2$ do
        for $n \in r$ do
            $p_m \leftarrow p_m * p(x_{i,j,n}^{(\mathtt{n})} \mid S_{m,n}^{(\mathtt{n})})$
        end for
    end for
end for
for $m \leftarrow 1$ to $\nu^{(\mathtt{d})} + 2$ do
    $p_m \leftarrow p_m * p(I_i^{(\mathtt{n})} = m \mid \mathcal{F}, \mathcal{V})$
end for
Sample $I_i^{(\mathtt{n})}$ with probability in proportion to $\boldsymbol{p}$
return $I_i^{(\mathtt{n})}$
```

5.4.4 Gibbs Sampling of Normal Indicator Variables $\mathcal{I}^{(\mathtt{n})}$

The conditional distribution of normal indicator variables is given by

$$\begin{aligned} & p(\mathcal{I}^{(\mathtt{n})} \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}, G, \mathcal{V}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})}) \\ & \quad \propto p(\mathcal{X}^{(\mathtt{n})} \mid \mathcal{S}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{n})}) p(\mathcal{I}^{(\mathtt{n})} \mid \mathcal{F}, \mathcal{V}) \\ & \quad = \prod_i \left(\prod_j \prod_n p(x_{i,j,n}^{(\mathtt{n})} \mid S_{I_i^{(\mathtt{n})},n}^{(\mathtt{n})}) \right) p(I_i^{(\mathtt{n})} \mid \mathcal{F}, \mathcal{V}). \end{aligned}$$

Therefore, each normal indicator variable $I_i^{(\mathtt{n})}$ can be sampled independently from the probability distribution in proportion to

$$\left(\prod_j \prod_n p(x_{i,j,n}^{(\mathtt{n})} \mid S_{I_i^{(\mathtt{n})},n}) \right) p(I_i^{(\mathtt{n})} \mid \mathcal{F}, \mathcal{V}),$$

using Gibbs sampling. The Gibbs sampling of each normal indicator variable is shown in Algorithm 5.4.

Algorithm 5.5 Gibbs sampling of each tumor indicator variable in ALPHLARD-NT

```
Input:
    $\mathcal{S}$: HLA sequences
    $\mathcal{F}$: likelihoods of indicator variables
    $G$: the ratio of normal cells contained in the tumor sample
    $\mathcal{V}$: validity flags
    $x_i^{(\mathtt{t})}$: the $i^{\mathrm{th}}$ tumor realigned read pair
Output:
    $I_i^{(\mathtt{t})}$: the $i^{\mathrm{th}}$ tumor indicator variable

$J \leftarrow \begin{cases} 1 & (\text{if } x_i^{(\mathtt{t})} \text{ is a paired read}) \\ 2 & (\text{if } x_i^{(\mathtt{t})} \text{ is an unpaired read}) \end{cases}$
for $m \leftarrow 1$ to $2\nu^{(\mathtt{d})} + 4$ do
    $p_m \leftarrow 1$
end for
for $j \leftarrow 1$ to $J$ do
    $r \leftarrow$ a set of positions covered by the read $x_{i,j}^{(\mathtt{t})}$
    for $m \leftarrow 1$ to $2\nu^{(\mathtt{d})} + 4$ do
        for $n \in r$ do
            $p_m \leftarrow p_m * p(x_{i,j,n}^{(\mathtt{t})} \mid S_{m,n})$
        end for
    end for
end for
for $m \leftarrow 1$ to $2\nu^{(\mathtt{d})} + 4$ do
    $p_m \leftarrow p_m * p(I_i^{(\mathtt{t})} = m \mid \mathcal{F}, G, \mathcal{V})$
end for
Sample $I_i^{(\mathtt{t})}$ with probability in proportion to $\boldsymbol{p}$
return $I_i^{(\mathtt{t})}$
```

5.4.5 Gibbs Sampling of Tumor Indicator Variables $\mathcal{I}^{(\mathtt{t})}$

The conditional distribution of tumor indicator variables is given by

$$\begin{aligned} &p(\mathcal{I}^{(\mathtt{t})} \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}, G, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})}) \\ &\quad \propto p(\mathcal{X}^{(\mathtt{t})} \mid \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{t})}) p(\mathcal{I}^{(\mathtt{t})} \mid \mathcal{F}, G, \mathcal{V}) \\ &\quad = \prod_i \left(\prod_j \prod_n p(x_{i,j,n}^{(\mathtt{t})} \mid S_{I_i^{(\mathtt{t})},n}) \right) p(I_i^{(\mathtt{t})} \mid \mathcal{F}, G, \mathcal{V}). \end{aligned}$$

Therefore, each tumor indicator variable $I_i^{(\mathtt{t})}$ can be sampled independently from the probability distribution in proportion to

$$\left(\prod_j \prod_n p(x_{i,j,n}^{(\mathtt{t})} \mid S_{I_i^{(\mathtt{t})},n}) \right) p(I_i^{(\mathtt{t})} \mid \mathcal{F}, G, \mathcal{V}),$$

using Gibbs sampling. The Gibbs sampling of each tumor indicator variable is shown in Algorithm 5.5.

5.4.6 Metropolis-Hastings Algorithm for Likelihoods $\mathcal{F}$ of Indicator Variables

For each F_m, a candidate parameter F_m^* is first sampled using the Metropolis-Hastings algorithm whose proposal distribution is given by

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

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

$$\begin{aligned} r &= \min(1, r^*), \\ r^* &= \frac{p(F_m^* \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}_{-m}, G, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})}{p(F_m \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}_{-m}, G, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})} \\ &\qquad \times \frac{p(F_m^* \to F_m \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}_{-m}, G, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})}{p(F_m, \to F_m^* \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}_{-m}, G, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})}. \end{aligned}$$

Each term can be obtained by

$$\begin{aligned} &p(F_m \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}_{-m}, G, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})}) \\ &\qquad \propto p(\mathcal{I}^{(\mathtt{n})} \mid F_m, \mathcal{F}_{-m}, \mathcal{V}) p(\mathcal{I}^{(\mathtt{t})} \mid F_m, \mathcal{F}_{-m}, G, \mathcal{V}) p(F_m), \\ &p(F_m, \to F_m^* \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}_{-m}, G, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})}) \\ &\qquad \propto 1. \end{aligned}$$

As a result,

$$r^* = \frac{p(\mathcal{I}^{(\mathtt{n})} \mid F_m^*, \mathcal{F}_{-m}, \mathcal{V})}{p(\mathcal{I}^{(\mathtt{n})} \mid F_m, \mathcal{F}_{-m}, \mathcal{V})} \frac{p(\mathcal{I}^{(\mathtt{t})} \mid F_m^*, \mathcal{F}_{-m}, G, \mathcal{V})}{p(\mathcal{I}^{(\mathtt{t})} \mid F_m, \mathcal{F}_{-m}, G, \mathcal{V})} \frac{p(F_m^*)}{p(F_m)}.$$

5.4.7 Metropolis-Hastings Algorithm for the Ratio G of Normal Cells Contained in the Tumor Sample

A candidate parameter G^* is first sampled using the Metropolis-Hastings algorithm whose proposal distribution is given by

$$G^* \sim \mathcal{LN}(G^* \mid \log G, (\sigma^{(\mathtt{g,p})})^2),$$

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

$$\begin{aligned} r &= \min(1, r^*) \\ r^* &= \frac{p(G^* \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})}{p(G \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})} \\ &\qquad \times \frac{p(G^* \to G \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})}{p(G, \to G^* \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})}. \end{aligned}$$

Each term can be obtained by

$$\begin{aligned} &p(G \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})}) \\ &\qquad \propto p(\mathcal{I}^{(\mathtt{t})} \mid \mathcal{F}, G, \mathcal{V}) p(G), \\ &p(G, \to G^* \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}, \mathcal{V}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})}) \\ &\qquad \propto 1. \end{aligned}$$

As a result,

$$r^* = \frac{p(\mathcal{I}^{(\mathtt{t})} \mid \mathcal{F}, G^*, \mathcal{V})}{p(\mathcal{I}^{(\mathtt{t})} \mid \mathcal{F}, G, \mathcal{V})} \frac{p(G^*)}{p(G)}.$$

5.4.8 Metropolis-Hastings Algorithm for Validity Flags $\mathcal{V}$

For each V_m, a candidate parameter V_m^* is sampled using Algorithm 5.6, whose proposal distribution is analogous to the Wolff algorithm [91], which is used for sampling of the Ising model. Then, $\mathcal{I}^{(\mathtt{n})*}$ and $\mathcal{I}^{(\mathtt{t})*}$ are also sampled using Gibbs sampling given V_m^*. The acceptance ratio r is calculated by

$$\begin{aligned}
r &= \min(1, r^*) \\
r^* &= \frac{p(V_m^*, \mathcal{I}^{(\mathtt{n})*}, \mathcal{I}^{(\mathtt{t})*} \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}, \mathcal{V}_{-m}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})}{p(V_m, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})} \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}, \mathcal{V}_{-m}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})}, \\
&\quad \times \frac{p(V_m^*, \mathcal{I}^{(\mathtt{n})*}, \mathcal{I}^{(\mathtt{t})*} \to V_m, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})} \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}, \mathcal{V}_{-m}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})}{p(V_m, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})} \to V_m^*, \mathcal{I}^{(\mathtt{n})*}, \mathcal{I}^{(\mathtt{t})*} \mid \mathcal{R}, \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, \mathcal{F}, \mathcal{V}_{-m}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})} \\
&= \frac{p(\mathcal{X}^{(\mathtt{n})} \mid \mathcal{S}^{(\mathtt{n})}, V_m^*, \mathcal{V}_{-m})}{p(\mathcal{X}^{(\mathtt{n})} \mid \mathcal{S}^{(\mathtt{n})}, V_m, \mathcal{V}_{-m})} \frac{p(\mathcal{X}^{(\mathtt{t})} \mid \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, V_m^*, \mathcal{V}_{-m})}{p(\mathcal{X}^{(\mathtt{t})} \mid \mathcal{S}^{(\mathtt{n})}, \mathcal{S}^{(\mathtt{t})}, V_m, \mathcal{V}_{-m})} \\
&\quad \times \frac{(\pi_v^{(\mathtt{v},\mathtt{p})})^{r-l}(1-\pi_v^{(\mathtt{v},\mathtt{p})})^{[l \neq 1 \wedge V_{l-1} \neq v] + [r \neq N \wedge V_{r+1} \neq v]}}{(\pi_{1-v}^{(\mathtt{v},\mathtt{p})})^{r-l}(1-\pi_{1-v}^{(\mathtt{v},\mathtt{p})})^{[l \neq 1 \wedge V_{l-1} \neq v] + [r \neq N \wedge V_{r+1} \neq v]}} \\
&\qquad \times \frac{\prod_{n=l}^{r+1} p(V_{m,n}^* \mid V_{m,n-1}^*)}{\prod_{n=l}^{r+1} p(V_{m,n} \mid V_{m,n-1})},
\end{aligned}$$

where v is a validity flag given in Algorithm 5.6. We set $1 - \pi^{(\mathtt{v},\mathtt{o})}$ and $\pi^{(\mathtt{v},\mathtt{e})}$ to $\pi_0^{(\mathtt{v},\mathtt{p})}$ and $\pi_1^{(\mathtt{v},\mathtt{p})}$, respectively, so that r^* can be calculated by

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

5.4.9 Metropolis-Hastings Algorithm for HLA Bases Not Covered with Reads

In addition to Gibbs sampling, we use the Metropolis-Hastings algorithm with a similar proposal distribution to that mentioned in Section 4.4.4, which considers HLA bases not covered with reads. First, for all $m \in \{1, \ldots, \nu^{(\mathtt{d})} + 2\}$, we define $S_m^{(\mathtt{n},\mathtt{N})}$ and $S_m^{(\mathtt{t},\mathtt{N})}$ by

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

Algorithm 5.6 Generate a candidate parameter V^* using the Wolff algorithm

```
Input:
    $V$: the current parameter
    $\pi_0^{(\mathtt{v,p})}$: probability for 0-cluster extension
    $\pi_1^{(\mathtt{v,p})}$: probability for 1-cluster extension
Output:
    $V^*$: candidate parameter

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

Here, $D_{m,n}^{(\mathtt{n})}$ is the number of normal realigned read pairs that cover $S_{m,n}$, $D_{m,n}^{(\mathtt{t})}$ is the number of tumor realigned read pairs that cover $S_{m,n}$, and $D_{m,n}$ is the number of realigned read pairs that cover $S_{m,n}^{(\mathtt{n})}$ or $S_{m,n}^{(\mathtt{t})}$. Thus, $S_m^{(\mathtt{n,N})}$ and $S_m^{(\mathtt{t,N})}$ are basically the same as $S_m^{(\mathtt{n})}$ and $S_m^{(\mathtt{t})}$, but bases that are not covered by any read are replaced with `N`'s. Then, A candidate HLA type R_m^*, a candidate normal HLA sequence $S_m^{(\mathtt{n})*}$, and a candidate tumor HLA sequence $S_m^{(\mathtt{t})*}$ are sampled by

$$\begin{aligned}
R_m^* &\sim p(R_m^* \mid S^{(\mathtt{n,N})}),\\
S_m^{(\mathtt{n})*} &\sim p(S_m^{(\mathtt{n})*} \mid R_m^*, \mathcal{S}_{-m}^{(\mathtt{n})}, S_m^{(\mathtt{t,N})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})}),\\
S_m^{(\mathtt{t})*} &\sim p(S_m^{(\mathtt{t})*} \mid S_m^{(\mathtt{n})*}, \mathcal{S}_{-m}^{(\mathtt{n})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{t})}).
\end{aligned}$$

The acceptance ratio r is given by

$$\begin{aligned}
r &= \min(1, r^*),\\
r^* &= \frac{p(R_m^*, S_m^{(\mathtt{n})*}, S_m^{(\mathtt{t})*} \mid \mathcal{R}_{-m}, \mathcal{S}_{-m}^{(\mathtt{n})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})}{p(R_m, S_m^{(\mathtt{n})}, S_m^{(\mathtt{t})} \mid \mathcal{R}_{-m}, \mathcal{S}_{-m}^{(\mathtt{n})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})}\\
&\quad \times \frac{p(R_m^*, S_m^{(\mathtt{n})*}, S_m^{(\mathtt{t})*} \to R_m, S_m^{(\mathtt{n})}, S_m^{(\mathtt{t})} \mid \mathcal{R}_{-m}, \mathcal{S}_{-m}^{(\mathtt{n})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})}{p(R_m, S_m^{(\mathtt{n})}, S_m^{(\mathtt{t})} \to R_m^*, S_m^{(\mathtt{n})*}, S_m^{(\mathtt{t})*} \mid \mathcal{R}_{-m}, \mathcal{S}_{-m}^{(\mathtt{n})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})}.
\end{aligned}$$

Each term can be obtained by

$$\begin{aligned}
&p(R_m, S_m^{(\mathtt{n})}, S_m^{(\mathtt{t})} \mid \mathcal{R}_{-m}, \mathcal{S}_{-m}^{(\mathtt{n})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})}) \\
&\qquad \propto p(\mathcal{X}^{(\mathtt{n})} \mid S_m^{(\mathtt{n})}, \mathcal{S}_{-m}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{n})}) p(\mathcal{X}^{(\mathtt{t})} \mid S_m^{(\mathtt{n})}, \mathcal{S}_{-m}^{(\mathtt{n})}, S_m^{(\mathtt{t})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{t})}) \\
&\qquad\qquad \times p(S_m^{(\mathtt{t})} \mid S_m^{(\mathtt{n})}) p(S_m^{(\mathtt{n})} \mid R_m) p(R_m), \\
&p(R_m, S_m^{(\mathtt{n})}, S_m^{(\mathtt{t})} \to R_m^*, S_m^{(\mathtt{n})*}, S_m^{(\mathtt{t})*} \mid \mathcal{R}_{-m}, \mathcal{S}_{-m}^{(\mathtt{n})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})}) \\
&\qquad = p(R_m^* \mid S^{(\mathtt{n},\mathtt{N})}) p(S_m^{(\mathtt{n})*} \mid R_m^*, \mathcal{S}_{-m}^{(\mathtt{n})}, S_m^{(\mathtt{t},\mathtt{N})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})}) \\
&\qquad\qquad \times p(S_m^{(\mathtt{t})*} \mid S_m^{(\mathtt{n})*}, \mathcal{S}_{-m}^{(\mathtt{n})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{t})}),
\end{aligned}$$

where

$$\begin{aligned}
&p(R_m^* \mid S^{(\mathtt{n},\mathtt{N})}) \\
&\qquad \propto p(S^{(\mathtt{n},\mathtt{N})} \mid R_m^*) p(R_m^*), \\
&p(S_m^{(\mathtt{n})*} \mid R_m^*, \mathcal{S}_{-m}^{(\mathtt{n})}, S_m^{(\mathtt{t},\mathtt{N})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})}) \\
&\qquad \propto \frac{p(\mathcal{X}^{(\mathtt{n})} \mid S_m^{(\mathtt{n})*}, \mathcal{S}_{-m}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{n})}) p(\mathcal{X}^{(\mathtt{t})} \mid S_m^{(\mathtt{n})*}, \mathcal{S}_{-m}^{(\mathtt{n})}, S_m^{(\mathtt{t},\mathtt{N})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{t})}) p(S_m^{(\mathtt{t},\mathtt{N})} \mid S_m^{(\mathtt{n})*}) p(S_m^{(\mathtt{n})*} \mid R_m^*)}{p(S_m^{(\mathtt{t},\mathtt{N})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})} \mid R_m^*, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})})}, \\
&p(S_m^{(\mathtt{t})*} \mid S_m^{(\mathtt{n})*}, \mathcal{S}_{-m}^{(\mathtt{n})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{t})}, \mathcal{X}^{(\mathtt{t})}) \\
&\qquad \propto \frac{p(\mathcal{X}^{(\mathtt{t})} \mid S_m^{(\mathtt{n})*}, \mathcal{S}_{-m}^{(\mathtt{n})}, S_m^{(\mathtt{t})*}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{t})}) p(S_m^{(\mathtt{t})*} \mid S_m^{(\mathtt{n})*})}{p(\mathcal{X}^{(\mathtt{t})} \mid S_m^{(\mathtt{n})*}, \mathcal{S}_{-m}^{(\mathtt{n})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{t})})}.
\end{aligned}$$

As a result,

$$\begin{aligned}
r^* = {} & \frac{p(S^{(\mathtt{n},\mathtt{N})} \mid R_m)}{p(S^{(\mathtt{n},\mathtt{N})} \mid R_m^*)} \frac{p(\mathcal{X}^{(\mathtt{t})} \mid S_m^{(\mathtt{n})}, \mathcal{S}_{-m}^{(\mathtt{n})}, S_m^{(\mathtt{t},\mathtt{N})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{t})})}{p(\mathcal{X}^{(\mathtt{t})} \mid S_m^{(\mathtt{n})*}, \mathcal{S}_{-m}^{(\mathtt{n})}, S_m^{(\mathtt{t},\mathtt{N})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{t})})} \frac{p(S_m^{(\mathtt{t},\mathtt{N})} \mid S_m^{(\mathtt{n})})}{p(S_m^{(\mathtt{t},\mathtt{N})} \mid S_m^{(\mathtt{n})*})} \\
& \times \frac{p(S_m^{(\mathtt{t},\mathtt{N})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})} \mid R_m^*, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})})}{p(S_m^{(\mathtt{t},\mathtt{N})}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})} \mid R_m, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})})} \frac{p(\mathcal{X}^{(\mathtt{t})} \mid S_m^{(\mathtt{n})*}, \mathcal{S}_{-m}^{(\mathtt{n})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{t})})}{p(\mathcal{X}^{(\mathtt{t})} \mid S_m^{(\mathtt{n})*}, \mathcal{S}_{-m}^{(\mathtt{n})}, \mathcal{S}_{-m}^{(\mathtt{t})}, \mathcal{I}^{(\mathtt{t})})}.
\end{aligned}$$

5.4.10 Metropolis-Hastings Algorithm for Swapping Non-Decoy Parameters and Decoy Parameters

In addition, we introduce the Metropolis-Hastings algorithm with a similar proposal distribution to that mentioned in Section 4.4.5, which swaps non-decoy parameters and decoy parameters to determine which HLA types and HLA sequences should be decoy parameters. However, we must modify the proposal distribution because ALPHLARD-NT has new parameters $\mathcal{V}$, which are not introduced in ALPHLARD. First, a non-decoy index m and a decoy index m' are uniformly sampled, that is,

$$\begin{aligned}
m &\sim \mathcal{U}(m \mid 1, 2), \\
m' &\sim \mathcal{U}(m' \mid 3, \nu^{(\mathtt{d})} + 2).
\end{aligned}$$

We also uniformly sample an interval i such that $\forall n \in i.V_{m',n} = 1$. This interval defines the swapped region; that is,

$$S^{(\mathtt{n})*}_{m,n} = \begin{cases} S^{(\mathtt{n})}_{m,n} & (\text{if } n \notin i) \\ S^{(\mathtt{n})}_{m',n} & (\text{if } n \in i) \end{cases},$$

$$S^{(\mathtt{n})*}_{m',n} = \begin{cases} S^{(\mathtt{n})}_{m',n} & (\text{if } n \notin i) \\ S^{(\mathtt{n})}_{m,n} & (\text{if } n \in i) \end{cases},$$

$$S^{(\mathtt{t})*}_{m,n} = \begin{cases} S^{(\mathtt{t})}_{m,n} & (\text{if } n \notin i) \\ S^{(\mathtt{t})}_{m',n} & (\text{if } n \in i) \end{cases},$$

$$S^{(\mathtt{t})*}_{m',n} = \begin{cases} S^{(\mathtt{t})}_{m',n} & (\text{if } n \notin i) \\ S^{(\mathtt{t})}_{m,n} & (\text{if } n \in i) \end{cases}.$$

Then, R^*_m, $R^*_{m'}$, $\mathcal{I}^{(\mathtt{n})*}$, and $\mathcal{I}^{(\mathtt{t})*}$ are sampled using Gibbs sampling by

$$\begin{aligned} R^*_m &\sim p(R^*_m \mid S^{(\mathtt{n})*}_m), \\ R^*_{m'} &\sim p(R^*_{m'} \mid S^{(\mathtt{n})*}_{m'}), \\ \mathcal{I}^{(\mathtt{n})*} &\sim p(\mathcal{I}^{(\mathtt{n})*} \mid S^{(\mathtt{n})*}_m, S^{(\mathtt{n})*}_{m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, \mathcal{F}, \mathcal{V}, \mathcal{X}^{(\mathtt{n})}), \\ \mathcal{I}^{(\mathtt{t})*} &\sim p(\mathcal{I}^{(\mathtt{t})*} \mid S^{(\mathtt{n})*}_m, S^{(\mathtt{n})*}_{m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, S^{(\mathtt{t})*}_m, S^{(\mathtt{t})*}_{m'}, \mathcal{S}^{(\mathtt{t})}_{-m,m'}, \mathcal{F}, G, \mathcal{V}, \mathcal{X}^{(\mathtt{t})}). \end{aligned}$$

The acceptance ratio r is given by

$$\begin{aligned} r &= \min(1, r^*), \\ r^* &= \frac{p(\boldsymbol{\theta}^*_{m,m'} \mid \mathcal{R}_{-m,m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, \mathcal{S}^{(\mathtt{t})}_{-m,m'}, \mathcal{F}, G, \mathcal{V}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})}{p(\boldsymbol{\theta}_{m,m'} \mid \mathcal{R}_{-m,m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, \mathcal{S}^{(\mathtt{t})}_{-m,m'}, \mathcal{F}, G, \mathcal{V}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})} \\ &\quad \times \frac{p(\boldsymbol{\theta}^*_{m,m'} \to \boldsymbol{\theta}_{m,m'} \mid \mathcal{R}_{-m,m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, \mathcal{S}^{(\mathtt{t})}_{-m,m'}, \mathcal{F}, G, \mathcal{V}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})}{p(\boldsymbol{\theta}_{m,m'} \to \boldsymbol{\theta}^*_{m,m'} \mid \mathcal{R}_{-m,m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, \mathcal{S}^{(\mathtt{t})}_{-m,m'}, \mathcal{F}, G, \mathcal{V}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})}, \\ \boldsymbol{\theta}_{m,m'} &= (R_m, R_{m'}, S^{(\mathtt{n})}_m, S^{(\mathtt{n})}_{m'}, S^{(\mathtt{t})}_m, S^{(\mathtt{t})}_{m'}, \mathcal{I}^{(\mathtt{n})}, \mathcal{I}^{(\mathtt{t})}), \\ \boldsymbol{\theta}^*_{m,m'} &= (R^*_m, R^*_{m'}, S^{(\mathtt{n})*}_m, S^{(\mathtt{n})*}_{m'}, S^{(\mathtt{t})*}_m, S^{(\mathtt{t})*}_{m'}, \mathcal{I}^{(\mathtt{n})*}, \mathcal{I}^{(\mathtt{t})*}). \end{aligned}$$

Each term can be obtained by

$$\begin{aligned} &p(\boldsymbol{\theta}_{m,m'} \mid \mathcal{R}_{-m,m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, \mathcal{S}^{(\mathtt{t})}_{-m,m'}, \mathcal{F}, G, \mathcal{V}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})})) \\ &\quad \propto p(\mathcal{X}^{(\mathtt{n})} \mid S^{(\mathtt{n})}_m, S^{(\mathtt{n})}_{m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, \mathcal{I}^{(\mathtt{n})}) \\ &\qquad \times p(\mathcal{X}^{(\mathtt{t})} \mid S^{(\mathtt{n})}_m, S^{(\mathtt{n})}_{m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, S^{(\mathtt{t})}_m, S^{(\mathtt{t})}_{m'}, \mathcal{S}^{(\mathtt{t})}_{-m,m'}, \mathcal{I}^{(\mathtt{t})}) \\ &\qquad\quad \times p(S^{(\mathtt{t})}_m \mid S^{(\mathtt{n})}_m) p(S^{(\mathtt{t})}_{m'} \mid S^{(\mathtt{n})}_{m'}) p(S^{(\mathtt{n})}_m \mid R_m) p(S^{(\mathtt{n})}_{m'} \mid R_{m'}) p(R_m) \\ &\qquad\qquad \times p(\mathcal{I}^{(\mathtt{n})} \mid \mathcal{F}, \mathcal{V}) p(\mathcal{I}^{(\mathtt{t})} \mid \mathcal{F}, G, \mathcal{V}), \\ &p(\boldsymbol{\theta}_{m,m'} \to \boldsymbol{\theta}^*_{m,m'} \mid \mathcal{R}_{-m,m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, \mathcal{S}^{(\mathtt{t})}_{-m,m'}, \mathcal{F}, G, \mathcal{V}, \mathcal{X}^{(\mathtt{n})}, \mathcal{X}^{(\mathtt{t})}) \\ &\quad \propto p(R^*_m \mid S^{(\mathtt{n})*}_m) p(R^*_{m'} \mid S^{(\mathtt{n})*}_{m'}) \\ &\qquad \times p(\mathcal{I}^{(\mathtt{n})*} \mid S^{(\mathtt{n})*}_m, S^{(\mathtt{n})*}_{m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, \mathcal{F}, \mathcal{V}, \mathcal{X}^{(\mathtt{n})}) \\ &\qquad\quad \times p(\mathcal{I}^{(\mathtt{t})*} \mid S^{(\mathtt{n})*}_m, S^{(\mathtt{n})*}_{m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, S^{(\mathtt{t})*}_m, S^{(\mathtt{t})*}_{m'}, \mathcal{S}^{(\mathtt{t})}_{-m,m'}, \mathcal{F}, G, \mathcal{V}, \mathcal{X}^{(\mathtt{t})}), \end{aligned}$$

where

$$\begin{aligned}
&p(R^*_m \mid S^{(\mathtt{n})*}_m) \\
&\qquad = \frac{p(S^{(\mathtt{n})*}_m \mid R^*_m) p(R^*_m)}{p(S^{(\mathtt{n})*}_m)}, \\
&p(R^*_{m'} \mid S^{(\mathtt{n})*}_{m'}) \\
&\qquad \propto \frac{p(S^{(\mathtt{n})*}_{m'} \mid R^*_{m'})}{p(S^{(\mathtt{n})*}_{m'})}, \\
&p(\mathcal{I}^{(\mathtt{n})*} \mid S^{(\mathtt{n})*}_m, S^{(\mathtt{n})*}_{m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, \mathcal{F}, \mathcal{V}, \mathcal{X}^{(\mathtt{n})}) \\
&\qquad = \frac{p(\mathcal{X}^{(\mathtt{n})} \mid S^{(\mathtt{n})*}_m, S^{(\mathtt{n})*}_{m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, \mathcal{I}^{(\mathtt{n})*}) p(\mathcal{I}^{(\mathtt{n})*} \mid \mathcal{F}, \mathcal{V})}{p(\mathcal{X}^{(\mathtt{n})} \mid S^{(\mathtt{n})*}_m, S^{(\mathtt{n})*}_{m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, \mathcal{F}, \mathcal{V})}, \\
&p(\mathcal{I}^{(\mathtt{t})*} \mid S^{(\mathtt{n})*}_m, S^{(\mathtt{n})*}_{m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, S^{(\mathtt{t})*}_m, S^{(\mathtt{t})*}_{m'}, \mathcal{S}^{(\mathtt{t})}_{-m,m'}, \mathcal{F}, G, \mathcal{V}, \mathcal{X}^{(\mathtt{t})}) \\
&\qquad = \frac{p(\mathcal{X}^{(\mathtt{t})} \mid S^{(\mathtt{n})*}_m, S^{(\mathtt{n})*}_{m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, S^{(\mathtt{t})*}_m, S^{(\mathtt{t})*}_{m'}, \mathcal{S}^{(\mathtt{t})}_{-m,m'}, \mathcal{I}^{(\mathtt{t})*}) p(\mathcal{I}^{(\mathtt{t})*} \mid \mathcal{F}, G, \mathcal{V})}{p(\mathcal{X}^{(\mathtt{t})} \mid S^{(\mathtt{n})*}_m, S^{(\mathtt{n})*}_{m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, S^{(\mathtt{t})*}_m, S^{(\mathtt{t})*}_{m'}, \mathcal{S}^{(\mathtt{t})}_{-m,m'}, \mathcal{F}, G, \mathcal{V})}.
\end{aligned}$$

Consequently, the acceptance ratio r^* is given by

$$\begin{aligned}
r^* = \frac{p(S^*_m)}{p(S_m} \frac{p(S^*_{m'})}{p(S_{m'}} &\frac{p(\mathcal{X}^{(\mathtt{n})} \mid S^{(\mathtt{n})*}_m, S^{(\mathtt{n})*}_{m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, \mathcal{F}, \mathcal{V})}{p(\mathcal{X}^{(\mathtt{n})} \mid S^{(\mathtt{n})}_m, S^{(\mathtt{n})}_{m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, \mathcal{F}, \mathcal{V})} \\
\times &\frac{p(\mathcal{X}^{(\mathtt{t})} \mid S^{(\mathtt{n})*}_m, S^{(\mathtt{n})*}_{m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, S^{(\mathtt{t})*}_m, S^{(\mathtt{t})*}_{m'}, \mathcal{S}^{(\mathtt{t})}_{-m,m'}, \mathcal{F}, G, \mathcal{V})}{p(\mathcal{X}^{(\mathtt{t})} \mid S^{(\mathtt{n})}_m, S^{(\mathtt{n})}_{m'}, \mathcal{S}^{(\mathtt{n})}_{-m,m'}, S^{(\mathtt{t})}_m, S^{(\mathtt{t})}_{m'}, \mathcal{S}^{(\mathtt{t})}_{-m,m'}, \mathcal{F}, G, \mathcal{V})}.
\end{aligned}$$

5.4.11 Strategies in the Burn-in Period

Other strategies were further used in the burn-in period to obtain better parameters. Some of the approaches are the same as those described in 4.4.6: the multi-start strategy and copying HLA sequences. In addition, we also use another approach that sequence reads are assigned to decoy sequences if there are mismatches between the sequence reads and the reference sequences. This approach helps to reduce the incidence of false-positive mutations and retains only the mutations that seem true.

5.4.12 HLA Analysis from Sampled Parameters

HLA analysis is conducted based on the sampled parameters. HLA genotyping is performed by counting the numbers of sampled HLA types. HLA germline mutation calling can be done by finding different bases between sampled HLA types and normal HLA sequences. In addition, we can also identify HLA somatic mutations by finding different bases between sampled normal HLA sequences and tumor HLA sequences.

5.5 Experimental Results

In this section, we illustrate the capability of ALPHLARD-NT using WGS data and WES data. First, we demonstrate the performance of ALPHLARD-NT for HLA genotyping from WGS data compared with ALPHLARD and POLYSOLVER [78]. We also show that ALPHLARD-NT can identify HLA somatic mutations that cannot be detected by other methods. Next, we show the results of HLA mutation calling from WES data. Lastly, we also discuss the effectiveness of decoy parameters in ALPHLARD-NT.

In the following sections, we used the most sampled HLA genotype in the MCMC process as the candidate HLA genotype in ALPHLARD-NT and ALPHLARD.

5.5.1 HLA Genotyping from WGS Data

We first evaluated the accuracy of ALPHLARD-NT for HLA genotyping from the WGS dataset used in Section 4.5 For comparison, we applied ALPHLARD-NT, ALPHLARD, and POLYSOLVER [78] to the WGS data. The performance comparison is summarized in Tables 5.1 and 5.2. Table 5.1 shows how many HLA types were correctly determined, and Table 5.2 shows how many samples were fully correctly genotyped. 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.

Table 5.1: WGS-based HLA genotyping accuracy that indicates how many HLA types were correctly determined with ALPHLARD-NT, ALPHLARD, and POLYSOLVER. N/A indicates that the method does not support the HLA gene or the resolution.

		ALPHLARD-NT	ALPHLARD	POLYSOLVER
HLA-A	1st	**100% (50/50)**	**100% (50/50)**	**100% (50/50)**
	2nd	**100% (50/50)**	98.0% (49/50)	98.0% (49/50)
	3rd	**98.0% (49/50)**	**98.0% (49/50)**	90.0% (45/50)
HLA-B	1st	**100% (48/48)**	**100% (48/48)**	91.7% (44/48)
	2nd	**100% (48/48)**	**100% (48/48)**	85.4% (41/48)
	3rd	**97.9% (47/48)**	95.8% (46/48)	81.3% (39/48)
HLA-C	1st	**100% (50/50)**	**100% (50/50)**	**100% (50/50)**
	2nd	**100% (50/50)**	98.0% (49/50)	90.0% (45/50)
	3rd	**100% (50/50)**	98.0% (49/50)	86.0% (43/50)
HLA-DPA1	1st	**100% (24/24)**	**100% (24/24)**	N/A
	2nd	**100% (24/24)**	**100% (24/24)**	N/A
	3rd	**100% (24/24)**	**100% (24/24)**	N/A
HLA-DPB1	1st	**100% (22/22)**	**100% (22/22)**	N/A
	2nd	**100% (22/22)**	**100% (22/22)**	N/A
	3rd	**100% (22/22)**	**100% (22/22)**	N/A
HLA-DQA1	1st	**100% (24/24)**	**100% (24/24)**	N/A
	2nd	**95.8% (23/24)**	**95.8% (23/24)**	N/A
	3rd	**95.8% (23/24)**	**95.8% (23/24)**	N/A
HLA-DQB1	1st	**100% (18/18)**	**100% (18/18)**	N/A
	2nd	**94.4% (17/18)**	**94.4% (17/18)**	N/A
	3rd	**94.4% (17/18)**	**94.4% (17/18)**	N/A
HLA-DRB1	1st	**100% (24/24)**	**100% (24/24)**	N/A
	2nd	**100% (24/24)**	**100% (24/24)**	N/A
	3rd	**100% (24/24)**	**100% (24/24)**	N/A
Total	1st	**100% (260/260)**	**100% (260/260)**	97.3% (144/148)
	2nd	**99.2% (258/260)**	98.5% (256/260)	91.2% (135/148)
	3rd	**98.5% (256/260)**	97.7% (254/260)	85.8% (127/148)

Table 5.2: WGS-based HLA genotyping accuracy that indicates how many samples were fully correctly genotyped with ALPHLARD-NT, ALPHLARD, and POLYSOLVER. N/A indicates that the method does not support the HLA gene or the resolution.

		ALPHLARD-NT	ALPHLARD	POLYSOLVER
HLA-A	1st	**100% (25/25)**	**100% (25/25)**	**100% (25/25)**
	2nd	**100% (25/25)**	96.0% (24/25)	96.0% (24/25)
	3rd	**96.0% (24/25)**	**96.0% (24/25)**	80.0% (20/25
HLA-B	1st	**100% (24/24)**	**100% (24/24)**	83.3% (20/24
	2nd	**100% (24/24)**	**100% (24/24)**	70.8% (17/24)
	3rd	**95.8% (23/24)**	91.7% (22/24)	62.5% (15/24)
HLA-C	1st	**100% (25/25)**	**100% (25/25)**	**100% (25/25)**
	2nd	**100% (25/25)**	96.0% (24/25)	80.0% (20/25)
	3rd	**100% (25/25)**	96.0% (24/25)	72.0% (18/25)
HLA-DPA1	1st	**100% (12/12)**	**100% (12/12)**	N/A
	2nd	**100% (12/12)**	**100% (12/12)**	N/A
	3rd	**100% (12/12)**	**100% (12/12)**	N/A
HLA-DPB1	1st	**100% (11/11)**	**100% (11/11)**	N/A
	2nd	**100% (11/11)**	**100% (11/11)**	N/A
	3rd	**100% (11/11)**	**100% (11/11)**	N/A
HLA-DQA1	1st	**100% (12/12)**	**100% (12/12)**	N/A
	2nd	**91.7% (11/12)**	**91.7% (11/12)**	N/A
	3rd	**91.7% (11/12)**	**91.7% (11/12)**	N/A
HLA-DQB1	1st	**100% (9/9)**	**100% (9/9)**	N/A
	2nd	**88.9% (8/9)**	**88.9% (8/9)**	N/A
	3rd	**88.9% (8/9)**	**88.9% (8/9)**	N/A
HLA-DRB1	1st	**100% (12/12)**	**100% (12/12)**	N/A
	2nd	**100% (12/12)**	**100% (12/12)**	N/A
	3rd	**100% (12/12)**	**100% (12/12)**	N/A
Total	1st	**100% (130/130)**	**100% (130/130)**	94.6% (70/74)
	2nd	**98.5% (128/130)**	96.9% (126/130)	82.4% (61/74)
	3rd	**96.9% (126/130)**	95.4% (124/130)	71.6% (53/74)

5.5.2 Detection of HLA Mutations from WGS 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 [77], 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 [90]. All of the insertions and the deletions are known to lead to the loss of function of the HLA types, and might contribute to immune escape. On the other hand, POLYSOLVER and EBCall detected no and one mutation, respectively, which would be likely due to the shallowness of the WGS data. In addition, ALPHLARD-NT is considered to be more sensitive than ALPHLARD because ALPHLARD can identify only two deletions and one insertion.

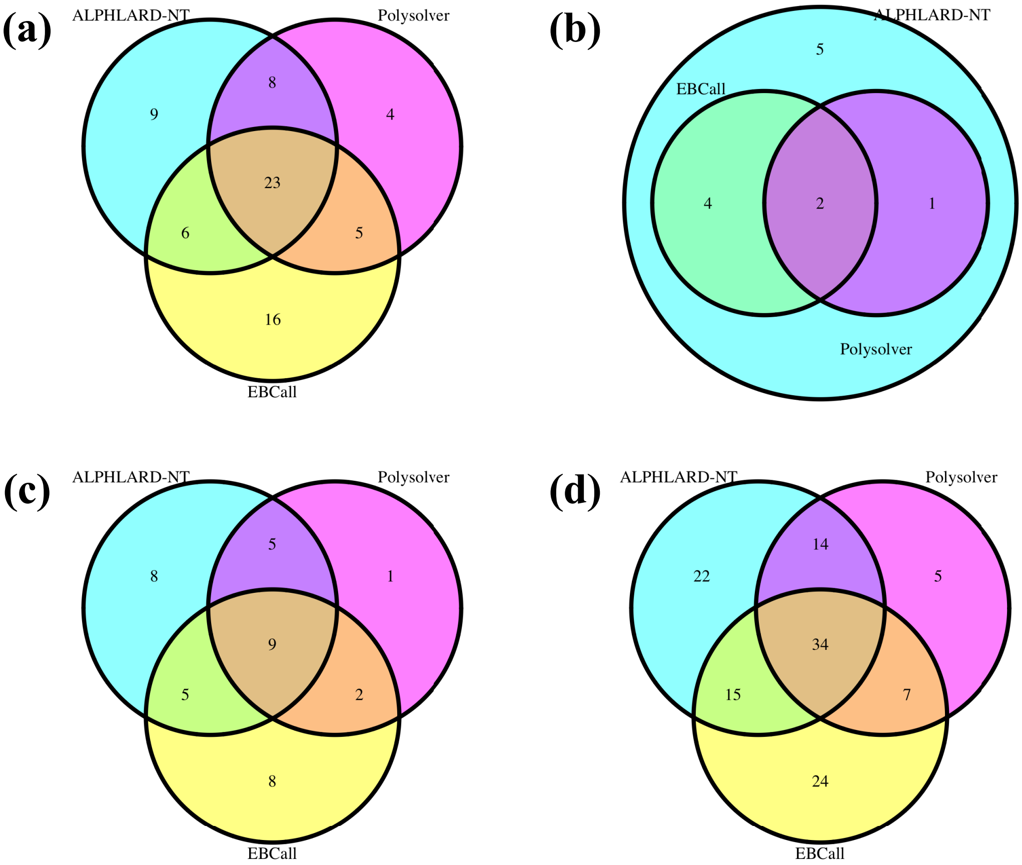


Figure 5.2: Venn diagrams of HLA somatic mutations identified by ALPHLARD-NT, POLYSOLVER, and EBCall for (a) substitutions, (b) insertions, (c) deletions, and (d) all mutations.

5.5.3 Detection of HLA Mutations from WES Data

Next, we applied ALPHLARD-NT, POLYSOLVER, and EBCall to a WES dataset of 343 colon adenocarcinoma cases from The Cancer Genome Atlas. Figure 5.2 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 to POLYSOLVER (60 mutations) and 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 insertions and deletions [60], 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 insertions and deletions seemed to be positively selected for immune escape caused by loss of function of the HLA types.

In addition, ALPHLARD-NT detected a novel HLA-B type 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 HLA type is also novel and not registered in the IPD-IMGT/HLA Database, indicating that the HLA type defines a new HLA type name at second field resolution.

5.5.4 Effectiveness of Decoy Parameters

We next demonstrate the effectiveness of decoy parameters in ALPHLARD-NT. We applied ALPHLARD-NT with decoy parameters and ALPHLARD-NT without decoy parameters to the WGS data. Tables 5.3 and 5.4 show the performance of the two approaches. These tables illustrate that the performance of ALPHLARD-NT with decoy parameters is slightly higher than that of ALPHLARD-NT without decoy parameters for the WGS data, but the difference is not significant. However, ALPHLARD-NT without decoy parameters failed to accurately identify HLA somatic mutations.

Figure 5.3 shows the Venn diagrams of the identified HLA class I somatic mutations from the WGS data using the two approaches. This figure indicates that the two versions identify the same insertions and deletions. However, ALPHLARD without decoy parameters call 20 additional substitutions, and we verified from the TruSight HLA Sequencing Panels that all of them are false-positive mutations. This suggests that ALPHLARD-NT without decoy parameters can accurately detect insertions and deletions but cannot identify substitutions. This is because the influence of misclassified reads could not be ignorant without decoy parameters.

Table 5.3: WGS-based HLA genotyping accuracy that indicates how many HLA types were correctly determined by ALPHLARD-NT with decoy parameters and without decoy parameters.

		with decoy	without decoy
HLA-A	1st	**100% (50/50)**	**100% (50/50)**
	2nd	**100% (50/50)**	**100% (50/50)**
	3rd	**98.0% (49/50)**	**98.0% (49/50)**
HLA-B	1st	**100% (48/48)**	**100% (48/48)**
	2nd	**100% (48/48)**	**100% (48/48)**
	3rd	**97.9% (47/48)**	95.8% (46/48)
HLA-C	1st	**100% (50/50)**	**100% (50/50)**
	2nd	**100% (50/50)**	**100% (50/50)**
	3rd	**100% (50/50)**	**100% (50/50)**
HLA-DPA1	1st	**100% (24/24)**	**100% (24/24)**
	2nd	**100% (24/24)**	**100% (24/24)**
	3rd	**100% (24/24)**	**100% (24/24)**
HLA-DPB1	1st	**100% (22/22)**	**100% (22/22)**
	2nd	**100% (22/22)**	**100% (22/22)**
	3rd	**100% (22/22)**	**100% (22/22)**
HLA-DQA1	1st	**100% (24/24)**	**100% (24/24)**
	2nd	**95.8% (23/24)**	91.7% (22/24)
	3rd	**95.8% (23/24)**	91.7% (22/24)
HLA-DQB1	1st	**100% (18/18)**	**100% (18/18)**
	2nd	94.4% (17/18)	**100% (18/18)**
	3rd	94.4% (17/18)	**100% (18/18)**
HLA-DRB1	1st	**100% (24/24)**	**100% (24/24)**
	2nd	**100% (24/24)**	95.8% (23/24)
	3rd	**100% (24/24)**	95.8% (23/24)
Total	1st	**100% (260/260)**	**100% (260/260)**
	2nd	**99.2% (258/260)**	98.8% (257/260)
	3rd	**98.5% (256/260)**	97.7% (254/260)

Table 5.4: WGS-based HLA genotyping accuracy that indicates how many samples were fully correctly genotyped by ALPHLARD-NT with decoy parameters and without decoy parameters.

		with decoy	without decoy
HLA-A	1st	**100% (25/25)**	**100% (25/25)**
	2nd	**100% (25/25)**	**100% (25/25)**
	3rd	**96.0% (24/25)**	**96.0% (24/25)**
HLA-B	1st	**100% (24/24)**	**100% (24/24)**
	2nd	**100% (24/24)**	**100% (24/24)**
	3rd	**95.8% (23/24)**	91.7% (22/24)
HLA-C	1st	**100% (25/25)**	**100% (25/25)**
	2nd	**100% (25/25)**	**100% (25/25)**
	3rd	**100% (25/25)**	**100% (25/25)**
HLA-DPA1	1st	**100% (12/12)**	**100% (12/12)**
	2nd	**100% (12/12)**	**100% (12/12)**
	3rd	**100% (12/12)**	**100% (12/12)**
HLA-DPB1	1st	**100% (11/11)**	**100% (11/11)**
	2nd	**100% (11/11)**	**100% (11/11)**
	3rd	**100% (11/11)**	**100% (11/11)**
HLA-DQA1	1st	**100% (12/12)**	**100% (12/12)**
	2nd	**91.7% (11/12)**	83.3% (10/12)
	3rd	**91.7% (11/12)**	83.3% (10/12)
HLA-DQB1	1st	**100% (9/9)**	**100% (9/9)**
	2nd	88.9% (8/9)	**100% (9/9)**
	3rd	88.9% (8/9)	**100% (9/9)**
HLA-DRB1	1st	**100% (12/12)**	**100% (12/12)**
	2nd	**100% (12/12)**	91.7% (11/12)
	3rd	**100% (12/12)**	91.7% (11/12)
Total	1st	**100% (130/130)**	**100% (130/130)**
	2nd	**98.5% (128/130)**	97.7% (127/130)
	3rd	**96.9% (126/130)**	95.4% (124/130)

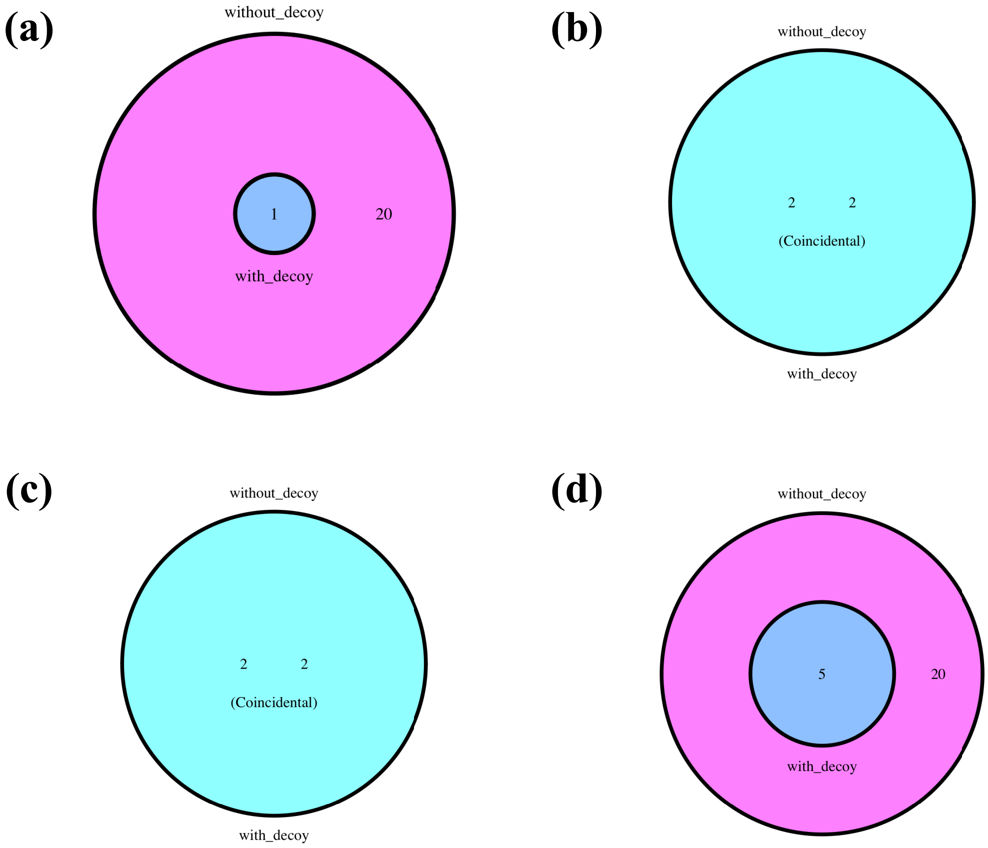


Figure 5.3: Venn diagrams of HLA somatic mutations identified by ALPHLARD-NT with decoy parameters and without decoy parameters for (a) substitutions, (b) insertions, (c) deletions, and (d) all mutations.

Chapter 6

Conclusion

6.1 Summary

Advances in next generation sequencing (NGS) technologies have promoted the research that focuses on the relationship between cancer and the immune system. Accordingly, human leukocyte antigen (HLA) analysis, including HLA genotyping, HLA germline mutation calling, and HLA somatic mutation calling, from NGS data has become essential to the research field. However, it is difficult to accurately analyze HLA genes from NGS data, such as whole exome sequence (WES), whole genome sequence (WGS), and RNA sequence (RNA-seq) data, mainly because HLA genes and HLA pseudogenes are similar to each other. Therefore, although it is relatively easy to extract reads that are produced by HLA genes and HLA pseudogenes, it is difficult to judge which HLA gene or HLA pseudogene produced each read. Especially, this problem is crucial when we conduct HLA analysis from WGS data due to the shallowness. Hence, we tackled the problem in this thesis.

In Chapter 4, we introduced an alignment-based scoring method to extract and classify sequence reads from HLA genes. The method begins with alignment of all sequence reads to all HLA types. We used the IPD-IMGT/HLA Database [69] to construct the reference sequences of HLA types. For each read and each HLA type, the HLA read score (HR score) is calculated. A read is classified into an HLA gene (i) if the maximum HR score in the HLA gene is sufficiently high and (ii) if the difference of the maximum scores between the HLA gene and the others is sufficiently large. The first condition (i) means that the read is likely to be produce by the HLA gene. On the other hand, the second condition (ii) means that the read is unlikely to be produced by the other HLA genes and HLA pseudogene than the HLA gene. The second condition is necessary because we should exclude sequence reads that are really produced by HLA genes and HLA pseudogenes that are similar to the HLA gene of interest. After extraction and classification of HLA reads, each read is realigned to the HLA type that achieved the highest HR score.

In Chapter 4, we presented a new Bayesian method, ALPHLARD, which performs accurate HLA genotyping from the realigned reads. ALPHLARD incorporates not only parameters for HLA types but also parameters for HLA sequences, which make it possible to call HLA germline mutations as well. ALPHLARD estimates the HLA genotype and the HLA germline mutations by calculating the posterior distribution using the Markov chain Monte Carlo (MCMC) method. The experimental results showed that ALPHLARD achieved higher accuracy for HLA genotyping from both WES and WGS data than existing methods. We presume that the high performance of ALPHLARD originates from the following reasons. First, the search space of ALPHLARD is all pairs of possible HLA types. Some methods treat an HLA genotype as two independent HLA types; that is, they give a score to each HLA type and output the most and the second most probable HLA types without directly considering the combinations. This approximation can reduce the computation time but works well only when the sequence data is sufficiently deep. Therefore, such methods would not achieve high accuracy for HLA genotyping from WGS data. Second, ALPHLARD uses decoy parameters in addition to non-decoy parameters. Hence, ALPHLARD can robustly and accurately perform HLA genotyping even if there exist some misclassified reads that are really produced by other HLA genes and HLA pseudogenes than the HLA gene of interest. We demonstrated that decoy parameters would enhance the performance of HLA genotyping.

In Chapter 5, we proposed a method, called ALPHLARD-NT, to conduct HLA somatic mutation calling as well as HLA genotyping and HLA germline mutation calling through simultaneous analysis of both normal and tumor sequence data of cancer patients. The statistical model of ALPHLARD-NT is obtained by extending ALPHLARD to include additional parameters for tumor sequence data. ALPHLARD-NT also contains parameters that control the ratio of sequence reads that are produced by each HLA sequence. As with ALPHLARD, ALPHLARD-NT also uses MCMC to estimate the posterior distribution of the parameters. The experimental results demonstrate that ALPHLARD-NT achieved higher performance for HLA genotyping from WGS data than other methods when both normal and tumor sequence data are available. Also, ALPHLARD-NT identified five HLA class I somatic mutations from the WGS data, although existing methods detected at most one somatic mutation. This suggests that ALPHLARD-NT can sensitively call HLA somatic mutation even from shallow sequence data. In addition, ALPHLARD-NT detected more insertions and deletions than the existing methods at a region which is known as a mutation hotspot, which indicates that the insertions and deletions seem true. We also showed that decoy parameters were effective for HLA somatic mutation calling in that they reduced false-positive mutations.

6.2 Future Work

6.2.1 Somatic Mutation Calling in HLA Class II Genes

Although we demonstrated that ALPHLARD-NT can identify somatic mutations in HLA class I genes, somatic mutation calling in HLA class II genes is still challenging because the IPD-IMGT/HLA database [69] is relatively incomplete for HLA class II pseudogenes, which leads to a large number of misclassified reads in analysis of HLA class II genes. One solution to this problem is identifying HLA sequences of HLA class II pseudogenes from a lot of samples using ALPHLARD-NT and registering the identified HLA sequences to the database.

6.2.2 Identification of Loss of Heterozygosity in HLA Genes

In addition to point mutations, there is another type of somatic mutations, loss of heterozygosity (LOH), which causes loss of a region in a chromosome. LOHs in HLA genes give a significant impact on the immune system because lost HLA types cannot work as components of the immune system. In addition, since the range of an LOH event is generally wide enough to cover all HLA genes, all of HLA types in a chromosome are lost if an LOH occurs in the HLA region. However, there is only one method to identify HLA LOHs, whose name is LOHHLA [56]. We would be able to construct a method to detect HLA LOHs by extending ALPHLARD-NT to include parameters for LOH events.

6.2.3 Calculation of Binding Affinity between an HLA Molecule and a Peptide

After HLA genotyping, it is important to calculate the binding affinities between identified HLA types and specific peptides such as viral peptides and mutated peptides to check which peptides are presented to T cells. There are some methods to calculate the binding affinity of an HLA type and a peptide [2, 35, 64]. These methods are based on neural network, but the structures are quite simple. Therefore, there remains room to improve by using more complicated structure such as long short-term memory [30] and residual connection [29].

References

[1] Cornelis A. Albers, Gerton Lunter, Daniel G. MacArthur, Gilean McVean, Willem H. Ouwehand, and Richard Durbin. Dindel: Accurate indel calls from short-read data. *Genome Research*, 21(6):961–973, 2011.

[2] Massimo Andreatta and Morten Nielsen. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics*, 32(4):511–517, 2015.

[3] Yu Bai, Min Ni, Blerta Cooper, Yi Wei, and Wen Fury. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. *BMC Genomics*, 15:325, 2014.

[4] Denis C. Bauer, Armella Zadoorian, Laurence O.W. Wilson, Melbourne Genomics Health Alliance, and Natalie P. Thorne. Evaluation of computational programs to predict HLA genotypes from genomic sequencing data. *Briefings in Bioinformatics*, 19(2):179–187, 2016.

[5] Sebastian Boegel, Martin Löwer, Michael Schäfer, Thomas Bukur, Jos De Graaf, Valesca Boisguérin, Özlem Türeci, Mustafa Diken, John C. Castle, and Ugur Sahin. HLA typing from RNA-Seq sequence reads. *Genome Medicine*, 4(12):102, 2012.

[6] Michael Burrows and David J. Wheeler. A Block-sorting Lossless Data Compression Algorithm. *Technical Report 124*, Palo Alto, CA: Digital Equipment Corporation, 1994.

[7] Andrea Cassinotti, Sarah Birindelli, Mario Clerici, Daria Trabattoni, Marco Lazzaroni, Sandro Ardizzone, Riccardo Colombo, Edoardo Rossi, and Gabriele Bianchi Porro. HLA and Autoimmune Digestive Disease: A Clinically Oriented Review for Gastroenterologists. *The American Journal of Gastroenterology*, 104(1):195, 2009.

[8] Subhra Chaudhuri, Annaiah Cariappa, Mei Tang, Daphne Bell, Daniel A. Haber, Kurt J. Isselbacher, Dianne Finkelstein, David Forcione, and Shiv Pillai. Genetic susceptibility to breast cancer: HLA DQB*03032 and HLA DRB1*11 may represent protective alleles. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11451–11454, 2000.

[9] Kristian Cibulskis, Michael S. Lawrence, Scott L. Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S. Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3):213, 2013.

[10] D Comas, E Mateu, F Calafell, A Pérez-Lezaun, E Bosch, R Martínez-Arias, and J Bertranpetit. HLA class I and class II DNA typing and the origin of Basques. *Tissue antigens*, 51(1):30–40, 1998.

[11] Alexander T. Dilthey, Pierre-Antoine Gourraud, Alexander J. Mentzer, Nezih Cereb, Zamin Iqbal, and Gil McVean. High-Accuracy HLA Type Inference from Whole-Genome Sequencing Data Using Population Reference Graphs. *PLOS Computational Biology*, 12(10):e1005151, 2016.

[12] Jiarui Ding, Ali Bashashati, Andrew Roth, Arusha Oloumi, Kane Tse, Thomas Zeng, Gholamreza Haffari, Martin Hirst, Marco A. Marra, Anne Condon, Samuel Aparicio, and Sohrab P. Shah. Feature-based classifiers for somatic mutation detection in tumour–normal paired sequencing data. *Bioinformatics*, 28(2):167–175, 2012.

[13] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[14] H. A. Elsner, J. Drábek, V. Rebmann, Z. Ambruzova, H. Grosse-Wilde, and R. Blasczyk. Non-expression of HLA-B*5111N is caused by an insertion into the cytosine island at exon 4 creating a frameshift stop codon. *Tissue Antigens*, 57(4):369–372, 2001.

[15] Rachel L. Erlich, Xiaoming Jia, Scott Anderson, Eric Banks, Xiaojiang Gao, Mary Carrington, Namrata Gupta, Mark A. DePristo, Matthew R. Henn, Niall J. Lennon, and Paul I.W. de Bakker. Next-generation sequencing for HLA typing of class I loci. *BMC Genomics*, 12:42, 2011.

[16] Thomas F. Gajewski, Yuru Meng, and Helena Harlin. Immune Suppression in the Tumor Microenvironment. *Journal of Immunotherapy*, 29(3):233–240, 2006.

[17] Cristina García-Corona, Elisa Vega-Memije, Adalberto Mosqueda-Taylor, Jesús K. Yamamoto-Furusho, Alma A. Rodríguez-Carreón, Jorge A. Ruiz-Morales, Norma Salgado, and Julio Granados. Association of HLA-DR4 (DRB1*0404) With Human Papillomavirus Infection in Patients With Focal Epithelial Hyperplasia. *Archives of Dermatology*, 140(10):1227–1231, 2004.

[18] Stuart Geman and Donald Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

[19] Charles J. Geyer. Markov chain Monte Carlo maximum likelihood. *Computing science and statistics: Proceedings of the 23rd Symposium on the Interface*, Fairfax StationInterface Foundation:156–163, 1991.

[20] Marios Giannakis, Xinmeng Jasmine Mu, Sachet A. Shukla, Zhi Rong Qian, Ofir Cohen, Reiko Nishihara, Samira Bahl, Yin Cao, Ali Amin-Mansour, Mai Yamauchi, Yasutaka Sukawa, Chip Stewart, Mara Rosenberg, Kosuke Mima, Kentaro Inamura, Katsuhiko Nosho, Jonathan A. Nowak, Michael S. Lawrence, Edward L. Giovannucci, Andrew T. Chan, Kimmie Ng, Jeffrey A. Meyerhardt, Eliezer M. Van Allen, Gad Getz, Stacey B. Gabriel, Eric S. Lander, Catherine J. Wu, Charles S. Fuchs, Shuji Ogino, and Levi A. Garraway.

Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. *Cell Reports*, 15(4):857–865, 2016.

[21] Faviel F. González-Galarza, Louise Y.C. Takeshita, Eduardo J.M. Santos, Felicity Kempson, Maria Helena Thomaz Maia, Andrea Luciana Soares da Silva, André Luiz Teles e Silva, Gurpreet S. Ghattaoraya, Ana Alfirevic, Andrew R. Jones, and Derek Middleton. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Research*, 43(D1):D784–D788, 2015.

[22] Robert R. Graham, Ward Ortmann, Peter Rodine, Karl Espe, Carl Langefeld, Ethan Lange, Adrienne Williams, Stephanie Beck, Chieko Kyogoku, Kathy Moser, Patrick Gaffney, Peter K. Gregersen, Lindsey A. Criswell, John B. Harley, and Timothy W. Behrens. Specific combinations of HLA-DR2 and DR3 class II haplotypes contribute graded risk for disease susceptibility and autoantibodies in human SLE. *European Journal of Human Genetics*, 15(8):823, 2007.

[23] Gregory R. Grant, Michael H. Farkas, Angel D. Pizarro, Nicholas F. Lahens, Jonathan Schug, Brian P. Brunk, Christian J. Stoeckert, John B. Hogenesch, and Eric A. Pierce. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, 27(18):2518–2528, 2011.

[24] Peter K. Gregersen, Jack Silver, and Robert J. Winchester. THE SHARED EPITOPE HYPOTHESIS. An Approach to Understanding The Molecular Genetics of Susceptibility to Rheumatoid Arthritis. *Arthritis & Rheumatism*, 30(11):1205–1213, 1987.

[25] Sergei I. Grivennikov, Florian R Greten, and Michael Karin. Immunity, Inflammation, and Cancer. *Cell*, 140(6):883–899, 2010.

[26] W. Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[27] Shuto Hayashi, Takuya Moriyama, Rui Yamaguchi, Shinichi Mizuno, Mitsuhiro Komura, Satoru Miyano, Hidewaki Nakagawa, and Seiya Imoto. ALPHLARD-NT: Bayesian method for HLA genotyping and mutation calling through simultaneous analysis of normal and tumor whole-genome sequence data. *Journal of Computational Biology*, in press.

[28] Shuto Hayashi, Rui Yamaguchi, Shinichi Mizuno, Mitsuhiro Komura, Satoru Miyano, Hidewaki Nakagawa, and Seiya Imoto. ALPHLARD: a Bayesian method for analyzing HLA genes from whole genome sequence data. *BMC Genomics*, 19(1):790, 2018.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[30] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural computation*, 9(8):1735–1780, 1997.

[31] Kazuyoshi Hosomichi, Timothy A. Jinam, Shigeki Mitsunaga, Hirofumi Nakaoka, and Ituro Inoue. Phase-defined complete sequencing of the HLA genes by next-generation sequencing. *BMC Genomics*, 14:355, 2013.

[32] International Cancer Genome Consortium. International network of cancer genome projects. *Nature*, 464(7291):993, 2010.

[33] E. Yvonne Jones, Lars Fugger, Jack L. Strominger, and Christian Siebold. MHC class II proteins and disease: a structural perspective. *Nature Reviews Immunology*, 6(4):271–282, 2006.

[34] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine learning*, 37(2):183–233, 1999.

[35] Vanessa Jurtz, Sinu Paul, Massimo Andreatta, Paolo Marcatili, Bjoern Peters, and Morten Nielsen. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *The Journal of Immunology*, page ji1700893, 2017.

[36] Daehwan Kim, Ben Langmead, and Steven L. Salzberg. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357, 2015.

[37] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L. Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36, 2013.

[38] Hyunsung John Kim and Nader Pourmand. HLA Haplotyping from RNA-seq Data Using Hierarchical Read Weighting. *PLoS One*, 8(6):e67885, 2013.

[39] Daniel C. Koboldt, Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, 2012.

[40] Sebastian Kreiter, Mathias Vormehr, Niels Van de Roemer, Mustafa Diken, Martin Löwer, Jan Diekmann, Sebastian Boegel, Barbara Schrörs, Fulvia Vascotto, John C. Castle, Arbel D. Tadmor, Stephen P. Schoenberger, Christoph Huber, Özlem Türeci, and Ugur Sahin. Mutant MHC class II epitopes drive therapeutic immune responses to cancer. *Nature*, 520(7549):692, 2015.

[41] M. Laforet, N. Froelich, A. Parissiadis, B. Pfeiffer, A. Schell, B. Faller, M. L. Woehl-Jaegle, J. P. Cazenave, and M. M. Tongio. A nucleotide insertion in exon 4 is responsible for the absence of expression of an HLA-A*01 allele. *Tissue Antigens*, 50(4):347–350, 1997.

[42] Kerstin Lang, Frank Entschladen, Corinna Weidt, and Kurt S. Zaenker. Tumor immune escape mechanisms: impact of the neuroendocrine system. *Cancer Immunology, Immunotherapy*, 55(7):749–760, 2006.

[43] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357, 2012.

[44] David E. Larson, Christopher C. Harris, Ken Chen, Daniel C. Koboldt, Travis E. Abbott, David J. Dooling, Timothy J. Ley, Elaine R. Mardis, Richard K. Wilson, and Li Ding. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317, 2011.

[45] Heewook Lee and Carl Kingsford. Kourami: graph-guided assembly for novel human leukocyte antigen allele discovery. *Genome Biology*, 19(1):16, 2018.

[46] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[47] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[48] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–1858, 2008.

[49] Ruiqiang Li, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, 2009.

[50] Yang Liao, Gordon K Smyth, and Wei Shi. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, 41(10):e108, 2013.

[51] Chang Liu, Xiao Yang, Brian Duffy, Thalachallour Mohanakumar, Robi D. Mitra, Michael C. Zody, and John D. Pfeifer. ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. *Nucleic Acids Research*, 41(14):e142, 2013.

[52] David Luckey, Dikshya Bastakoty, and Ashutosh K. Mangalam. Role of HLA class II genes in susceptibility and resistance to multiple sclerosis: Studies using HLA transgenic mice. *Journal of Autoimmunity*, 37(2):122–128, 2011.

[53] Katharine E. Magor, Eleanor J. Taylor, Susan Y. Shen, Eduardo Martinez-Naves, Nicholas M. Valiante, R. Spencer Wells, Jenny E. Gumperz, Erin J. Adams, Ann-Margaret Little, Fionnuala Williams, D. Middleton, X. Gao, J. McCluskey, P. Parham, and K. Lienert-Weidenbach. Natural inactivation of a common HLA allele (A*2402) has occurred on at least three separate occasions. *The Journal of Immunology*, 158(11):5242–5250, 1997.

[54] Steven G. E. Marsh, E. D. Albert, W. F. Bodmer, R. E. Bontrop, B. Dupont, H. A. Erlich, M. Fernández-Viña, D. E. Geraghty, R. Holdsworth, C. K. Hurley, M. Lau, K. W. Lee, B. Mach, M. Maiers, W. R. Mayr, C. R. Müller, P. Parham, E. W. Petersdorf, T. Sasazuki, J. L. Strominger, A. Svejgaard, P. I. Terasaki, J. M. Tiercy, and J. Trowsdale. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens*, 75(4):291–455, 2010.

[55] Rachel Marty, Saghar Kaabinejadian, David Rossell, Michael J. Slifker, Joris van de Haar, Hatice Billur Engin, Nicola de Prisco, Trey Ideker, William H. Hildebrand, Joan Font-Burgada, and Hannah Carter. MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell*, 171(6):1272–1283, 2017.

[56] Nicholas McGranahan, Rachel Rosenthal, Crispin T. Hiley, Andrew J. Rowan, Thomas B.K. Watkins, Gareth A. Wilson, Nicolai J. Birkbak, Selvaraju Veeriah, Peter Van Loo, Javier Herrero, Charles Swanton, and the TRACERx Consortium. Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell*, 171(6):1259–1271, 2017.

[57] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

[58] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[59] Emmanuel Mignot. Genetics of Narcolepsy and Other Sleep Disorders. *The American Journal of Human Genetics*, 60(6):1289–1302, 1997.

[60] Shinichi Mizuno, Rui Yamaguchi, Takanori Hasegawa, Shuto Hayashi, Masashi Fujita, Fan Zhang, Youngil Koh, Su-Yeon Lee, Sung-Soo Yoon, Eigo Shimizu, Mitsuhiro Komura, Akihiro Fujimoto, Momoko Nagai, Mamoru Kato, Han Liang, Satoru Miyano, Zemin Zhang, Hidewaki Nakagawa, and Seiya Imoto. Immuno-genomic PanCancer Landscape Reveals Diverse Immune Escape Mechanisms and Immuno-Editing Histories. *bioRxiv*, page 285338, 2018.

[61] Ali Mortazavi, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621, 2008.

[62] Naoki Nariai, Kaname Kojima, Sakae Saito, Takahiro Mimori, Yukuto Sato, Yosuke Kawai, Yumi Yamaguchi-Kabata, Jun Yasuda, and Masao Nagasaki. HLA-VBSeq: accurate HLA typing at full resolution from whole-genome sequencing data. *BMC Genomics*, 16(2):S7, 2015.

[63] Janelle A. Noble and Ana M. Valdes. Genetics of the HLA Region in the Prediction of Type 1 Diabetes. *Current Diabetes Reports*, 11(6):533, 2011.

[64] Timothy J O'Donnell, Alex Rubinsteyn, Maria Bonsack, Angelika B Riemer, Uri Laserson, and Jeff Hammerbacher. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell systems*, 7(1):129–132, 2018.

[65] Shigeaki Ohno, Masaki Ohguchi, Shigeto Hirose, Hidehiko Matsuda, Akemi Wakisaka, and Miki Aizawa. Close Association of HLA-Bw51 With Behçet's Disease. *Archives of Ophthalmology*, 100(9):1455–1458, 1982.

[66] Ole Olerup and Henrik Zetterquist. 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(5):225–235, 1992.

[67] R. Prieto-Pérez, T. Cabaleiro, E. Daudén, and F. Abad-Santos. Gene polymorphisms that can predict response to anti-TNF therapy in patients with psoriasis and related autoimmune diseases. *The Pharmacogenomics Journal*, 13(4):297–305, 2013.

[68] Licia Rivoltini, Paola Canese, Veronica Huber, Manuela Iero, Lorenzo Pilla, Roberta Valenti, Stefano Fais, Francesco Lozupone, Chiara Casati, Chiara Castelli, and Giorgio Parmiani. Escape strategies and reasons for failure in the interaction between tumour cells and the immune system: how can we

tilt the balance towards immune-mediated cancer control? *Expert Opinion on Biological Therapy*, 5(4):463–476, 2005.

[69] James Robinson, Jason A. Halliwell, James D. Hayhurst, Paul Flicek, Peter Parham, and Steven G.E. Marsh. The IPD and IMGT/HLA Database: allele variant databases. *Nucleic Acids Research*, 43(D1):D423–D431, 2015.

[70] Michael S. Rooney, Sachet A. Shukla, Catherine J. Wu, Gad Getz, and Nir Hacohen. Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. *Cell*, 160(1-2):48–61, 2015.

[71] Andrew Roth, Jiarui Ding, Ryan Morin, Anamaria Crisan, Gavin Ha, Ryan Giuliany, Ali Bashashati, Martin Hirst, Gulisa Turashvili, Arusha Oloumi, Marco A. Marra, Samuel Aparicio, and Sohrab P. Shah. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics*, 28(7):907–913, 2012.

[72] Randall K. Saiki, Teodorica L. Bugawan, Glenn T. Horn, Kary B. Mullis, and Henry A. Erlich. Analysis of enzymatically amplified β-globin and HLA-DQα DNA with allele-specific oligonucleotide probes. *Nature*, 324(6093):163, 1986.

[73] Pere Santamaria, Michael T. Boyce-Jacino, Alan L. Lindstrom, Jose J. Barbosa, Anthony J. Faras, and Stephen S. Rich. HLA class II "typing": direct sequencing of DRB, DQB, and DQA genes. *Human Immunology*, 33(2):69–81, 1992.

[74] Christopher T. Saunders, Wendy S. W. Wong, Sajani Swamy, Jennifer Becq, Lisa J. Murray, and R. Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14):1811–1817, 2012.

[75] Lee Schlosstein, Paul I. Terasaki, Rodney Bluestone, and Carl M. Pearson. High Association of an HL-A Antigen, W27, with Ankylosing Spondylitis. *The New England Journal of Medicine*, 288(14):704–706, 1973.

[76] Robert D. Schreiber, Lloyd J. Old, and Mark J. Smyth. Cancer Immunoediting: Integrating Immunity's Roles in Cancer Suppression and Promotion. *Science*, 331(6024):1565–1570, 2011.

[77] Yuichi Shiraishi, Yusuke Sato, Kenichi Chiba, Yusuke Okuno, Yasunobu Nagata, Kenichi Yoshida, Norio Shiba, Yasuhide Hayashi, Haruki Kume, Yukio Homma, Masashi Sanada, Seishi Ogawa, and Satoru Miyano. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Research*, 41(7):e89, 2013.

[78] Sachet A. Shukla, Michael S. Rooney, Mohini Rajasagi, Grace Tiao, Philip M. Dixon, Michael S. Lawrence, Jonathan Stevens, William J. Lane, Jamie L. Dellagatta, Scott Steelman, Carrie Sougnez, Kristian Cibulskis, Adam Kiezun, Nir Hacohen, Vladimir Brusic, Catherine J. Wu, and Gad Getz. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nature Biotechnology*, 33(11):1152–1158, 2015.

[79] D. M. Smith, W. B. Gardner, J. E. Baker, S. T. Cox, and L. A. Kresie. A new HLA-A*31 null allele, A*3114N. *Tissue Antigens*, 68(6):526–527, 2006.

[80] Robert H. Swendsen and Jian-Sheng Wang. Replica Monte Carlo Simulation of Spin-Glasses. *Physical Review Letters*, 57(21):2607, 1986.

[81] András Szolek, Benjamin Schubert, Christopher Mohr, Marc Sturm, Magdalena Feldhahn, and Oliver Kohlbacher. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*, 30(23):3310–3316, 2014.

[82] M. Testoni, E. Zucca, K.H. Young, and F. Bertoni. Genetic lesions in diffuse large B-cell lymphomas. *Annals of Oncology*, 26(6):1069–1080, 2015.

[83] The Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517(7536):576, 2015.

[84] The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, 513(7517):202, 2014.

[85] The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299, 2005.

[86] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68, 2015.

[87] Kai Wang, Darshan Singh, Zheng Zeng, Stephen J. Coleman, Yan Huang, Gleb L. Savich, Xiaping He, Piotr Mieczkowski, Sara A. Grimm, Charles M. Perou, James N. MacLeod, Derek Y. Chiang, Jan F. Prins, and Jinze Liu. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, 38(18):e178, 2010.

[88] René L. Warren, Gina Choe, Douglas J. Freeman, Mauro Castellarin, Sarah Munro, Richard Moore, and Robert A. Holt. Derivation of HLA types from shotgun sequence datasets. *Genome Medicine*, 4(12):95, 2012.

[89] David Weese, Manuel Holtgrewe, and Knut Reinert. RazerS 3: Faster, fully sensitive read mapping. *Bioinformatics*, 28(20):2592–2599, 2012.

[90] Eric T Weimer, M Montgomery, R Petraroia, et al. Performance Characteristics and Validation of Next-Generation Sequencing for Human Leucocyte Antigen Typing. *J Mol Diagn*, 18(5):668–675, 2016.

[91] Ulli Wolff. Collective Monte Carlo Updating for Spin Systems. *Phys Rev Lett*, 62(4):361, 1989.

[92] Thomas D. Wu and Serban Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, 2010.

[93] Chao Xie, Zhen Xuan Yeo, Marie Wong, Jason Piper, Tao Long, Ewen F. Kirkness, William H. Biggs, Ken Bloom, Stephen Spellman, Cynthia Vierra-Green, Colleen Brady, Richard H. Scheuermann, Amalio Telenti, Sally Howard, Suzanne Brewerton, Yaron Turpaz, and J. Craig Venter. Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. *Proceedings of the National Academy of Sciences of the United States of America*, 114(30):8059–8064, 2017.