論文の内容の要旨

Abstract

論文題目　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）

氏　名　林　周斗

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.