Computational Approaches to Prediction and Analysis of Human Leukocyte Antigen Genes

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Vanja Paunić

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Advisor: Vipin Kumar
Co-advisor: Michael Steinbach

March, 2014

Acknowledgements

I am here today because many people helped me along the way.

First, I would like thank my advisors, Prof. Vipin Kumar and Prof. Michael Steinbach, for their continuous support and guidance and for letting me work on the projects I found interesting that have, at the end, resulted in this thesis. Thank you, I have learned so much from you!

I would like to thank Mr. Daniel Whalen for giving me an opportunity to study in the United States through a scholarship I received from the Whalen Family Foundation. His benevolent desire to do good in the world brought me here and opened a world of opportunities.

I would also like to thank Georganne Tolaas and Laura Connor from the UMN Computer Science department for their invaluable help navigating the logistics and requirements of the grad school. You made the years spent in the CS department easier and more enjoyable!

Finally, I would like to thank my family. My parents and my sister, although far away, celebrated even my smallest accomplishments. Their unconditional confidence in me helped me climb many obstacles, and their support and love made it easier when I'd fall. My husband Mike traveled this journey with me as a fellow graduate student, celebrated all my successes, encouraged me during setbacks, and made sure I never felt alone. I have my family to thank for everything I am and what I achieved in my life, including this thesis.

Dedication

Za moje roditelje, Daru i Pavla.

To my parents, Dara and Pavle.

Abstract

The Human Leukocyte Antigen (HLA) gene system is the most polymorphic region of the human genome, containing some of the strongest associations with autoimmune, infectious, and inflammatory diseases. It plays a crucial role in hematopoietic stem cell transplantation, where patients and donors are matched with respect to their HLA genes to maximize the chances of a successful transplant. As such, HLA data is a highly valuable asset for clinicians and researchers for elucidating various disease-driving biological mechanisms. This thesis contains original research on the analysis of uncertainty in HLA data, exploration of the strong correlation structure in the region and prediction of HLA genes from widely available genetic markers.

We start by describing a novel method for correlated multi-label, multi-class prediction, which aims to solve the problem of prediction of HLA genes from widely available Single Nucleotide Polymorphism (SNP) data. Direct typing of HLA genes for large studies is expensive due to their extreme genetic polymorphism. Therefore, obtaining the HLA genes by prediction, rather than genetic typing, would be highly time- and cost-effective. In this study we use a two-step approach, involving label (gene) independent classifiers and label dependencies in the form of HLA haplotype frequencies, to predict HLA genes from SNP data. In addition, we propose different ways of integrating label dependency information into the prediction process and evaluate their impact on the prediction performance. The results from experiments on real-world data sets show that adding label dependencies into the prediction of HLA genes increases prediction accuracy when compared against the gene-independent approach.

Next, we aim to resolve and quantify the uncertainty that exists in HLA data sets. Due to the high genetic polymorphism of HLA genes, their molecular typing often results in a set of uncertain or ambiguous assignments, rather than an exact allele assignment at each gene. We propose a novel, information theoretic measure to quantify uncertainty in HLA typing. In addition, we demonstrate that using the HLA gene dependencies that reflect the strong correlation structure in the region, decreases the uncertainty in HLA data.

In the fourth chapter of the thesis, we propose a novel approach for multi-label prediction from uncertain data in the context of SNP-based prediction of HLA genes using ambiguous HLA data in training. Most existing HLA data sets contain uncertainty and, as such, need to be imputed to exact data before being used for training prediction models. Existing approaches for prediction of HLA genes from SNP data do not accommodate learning from uncertain data and, as such, miss the potential for an increased sample size and consequently improvements in prediction performance. In this thesis, we propose a novel algorithm for SNP-based prediction of HLA genes that utilizes ambiguous HLA data for building the prediction model. Additionally, we measure the impact that the uncertainty in the training data has on the prediction accuracy, and evaluate it on a real world data set. Our results show that the prediction from ambiguous HLA data generally performs better than the alternative approach which first imputes the ambiguous data into high-resolution HLA alleles and uses it to build the model.

The work in this thesis is a step toward understanding the immense challenges in the analysis of the HLA gene system. In this thesis, we: *i)* define and solve a problem of prediction of HLA genes from widely available genetic markers using a correlated multi-label, multi-class approach, *ii)* define and validate a measure to quantify the uncertainty present in HLA data sets, and *iii)* propose a novel approach to correlated prediction from uncertain data in the context of prediction of HLA genes. We conclude the thesis by discussing future work to further the understanding of this important genetic region through novel computational algorithms.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

1.1.1 Human leukocyte antigen genes

The Human Leukocyte Antigen (HLA) gene system on chromosome 6, also known as the Major Histocompatibility Complex (MHC) region, is the most polymorphic region of the human genome and one of the most extensively studied regions due to its importance in transplantation and association with autoimmune, infectious and inflammatory diseases [14, 16]. The HLA region contains genes that encode proteins crucial for adaptive immune response, and its high genetic polymorphism allows the immune system to fight a variety of adversities. This gene system plays a critical role in hematopoietic stem cell transplantation (HSCT), where patients and donors are matched with respect to their HLA genes in order to maximize the chances of a successful transplant [19]. HLA genes that are most intensively studied are the so-called classical HLA genes divided into two regions: class I (HLA-A, -B, -C) and class II (HLA-DR, -DQ, -DP) as shown in Figure 1.1.

The HLA genes are highly polymorphic - there are many variations of these genes in a population. We refer to variations of the genes as *alleles*. Developments in DNA-based typing methods have seen a large increase in new HLA alleles being identified each year with an average rate of over one new allele discovered per day [38, 72]. As of today, more than $9,000$ alleles have been discovered for the class-I, HLA-A, -C, -B,

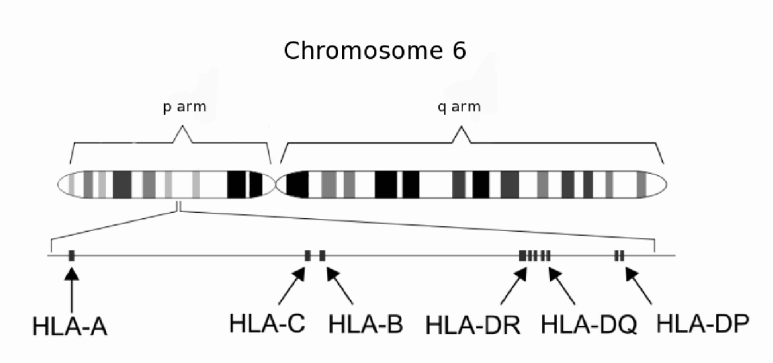

Figure 1.1: Human leukocyte antigen (HLA) gene system on chromosome 6. Shown are the genomic locations of 6 HLA genes, HLA-A, -C, -B, -DR, -DQ, -DP.

and class II, HLA-DRB1, -DQB1 genes [64]. The chance of two unrelated individuals having identical HLA alleles on all genes is very low since the diversity of HLA in the human population is one aspect of disease defense.

Naming of HLA alleles is standardized and regulated by the World Health Organization Nomenclature Committee for Factors of the HLA System [40]. Each allele starts with the gene name (e.g., A, B, DRB1, etc.), followed by at least two sets of digits, separated by colons (Figure 1.2). The first set of digits corresponds to the serological antigen encoded by the allele, while the second one corresponds to a specific subtype. For example, allele $A*01:02$ is found on gene A, belongs to 01 allele group and encodes a protein coded as 02. If necessary, longer names are assigned up to a total of four sets of digits. Two-digit resolution, also referred to as allele-family level, groups alleles into functionally related categories and is not protein-specific. In stem cell transplants, patients and donors are matched with respect to their four-digit alleles, since they identify the exact protein product of the gene. This resolution is commonly referred to as high-resolution.

1.1.2 Uncertainty (ambiguity) in HLA data

The high polymorphism in HLA presents a challenge when it comes to typing or sequencing HLA genes. The typing has historically been performed using serological antibody tests, which are able to identify HLA protein variants on the surface of the cell using

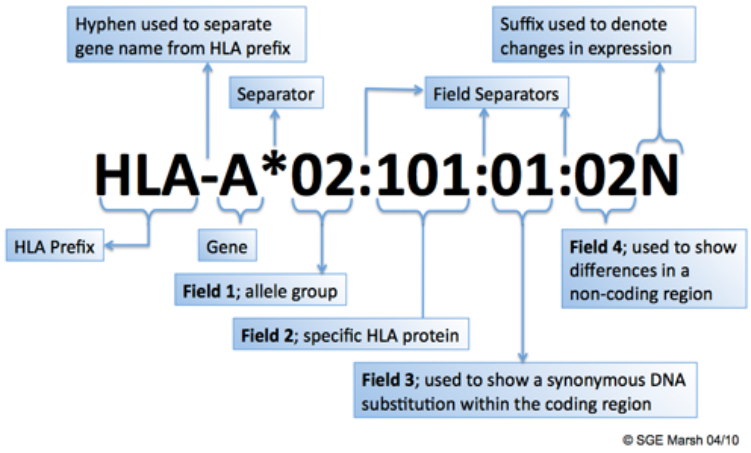

Figure 1.2: HLA allele naming [63]

antigen-specific antibodies [9]. Serology has been widely replaced with DNA-based typing methods due to its inability to identify all specific variations of the HLA alleles. It has been shown that in order to improve the clinical outcome of stem cell transplantation from an unrelated donor, it is essential to identify and match patient and donor's HLA genes at the allele level [75, 46, 57, 66].

DNA-based methods identify HLA alleles by interrogating the nuclear DNA sequence and can result in different levels of ambiguity depending on typing methodology or test kits used. HLA typing methods, their corresponding formats and abbreviations are shown in Table 3.1. Some of the widely used molecular methods for typing HLA genes, as defined in American Society for Histocompatibility and Immunogenetics Standards (ASHI), are a nucleic acid-based typing method using sequence-specific oligonucleotide hybridization (SSO) [25, 48], a nucleic acid amplification-based typing method using sequence-specific priming (SSP) and sequence-based typing (SBT) [65]. Even though DNA-based technology improved identification of specific alleles, HLA typing results reported by testing laboratories are still commonly resolved to a certain level of ambiguity, rather than to an exact allele assignment [23]. Exact high resolution HLA typing can be costly and laborious with the large and rapidly growing number of described HLA alleles, which sometimes cannot be easily distinguished with existing high-throughput typing methods. Ambiguous allele assignments are produced either due to failure to interrogate all polymorphic positions, or due to a lack of phase between polymorphisms

within a gene because of diploid sequence reads (or both).

In HSCT the selection of donors for a patient in need of a transplant is based primarily on HLA matching, and the lower the ambiguity of typing the easier it is to determine the probability of allele level match during the donor search [26]. In order to facilitate rapid identification of matched donors for HSCT several methods have been proposed to infer unknown phase and allele assignment [19, 35]. They typically employ statistical methods and the unique properties of the HLA region, such as its high linkage disequilibrium, in order to estimate haplotype frequencies and predict the most likely haplotype assignment for an individual with an HLA typing consisting of a set of ambiguous allele pairs. HLA typing with less ambiguity on average gives fewer high-probability phased high resolution haplotypes. Resolving the uncertainty in HLA data is important to all its relevant downstream uses.

1.2 Thesis Contributions

1.2.1 Correlated prediction of HLA genes from SNP data

Direct typing of HLA alleles for large studies is prohibitively expensive. It is estimated that typing of HLA alleles of class I and class II genes is about as expensive as typing 500,000 single nucleotide polymorphisms (SNPs), mostly bi-allelic genetic markers, across the genome [32]. However, HLA genes could be a causal or confounding factor in large studies due to their strong disease associations [14, 16] and some of the strongest adverse drug reactions [42, 41]. Previous studies have shown that single nucleotide polymorphisms and other genetic markers show strong associations with HLA alleles [11, 39]. This finding motivated work in the area of HLA prediction from SNP genotypes with the goal of enriching available large genome-wide studies with this valuable information. These methods have been successfully applied on SNP data sets to infer HLA genes and demonstrated the ability to confirm known drug and disease associations found in HLA-based studies using SNP data only [10, 42].

Initial attempts in HLA inference from SNP data used tag SNPs to infer common HLA alleles [11]. This study observed that there is a strong association (linkage disequilibrium (LD)) [13] between genetic markers and some HLA alleles and listed several such associations. While the study demonstrated efficient prediction of some common

HLA alleles using one or two SNP markers, this approach does not provide a solution for accurate prediction of all HLA alleles due to their high polymorphism and the existence of many rare alleles, which are hard to predict using a small number of tag SNPs.

More sophisticated methods have since been developed and used to predict HLA alleles from SNP data. These methods can be roughly divided into two categories: *i)* methods that use phased[1] SNP haplotypes to predict HLA alleles [32, 70, 43] and *ii)* methods that use un-phased SNP genotypes to predict HLA alleles [34, 12, 85, 27]. Empirical evaluation of known issues in HLA prediction is described in [84] and demonstrated with the methods in [34] on several datasets of different ancestries and different typing methodologies.

All existing approaches make predictions for HLA alleles for each gene independently. They consider each HLA gene as an independent label and build predictive models with respect to each label separately. Since many HLA alleles are known to be in strong linkage disequilibrium [13], we recently described an approach that takes advantage of this label dependency to assist the HLA inference. For example, we might predict an individual's HLA-C allele with high confidence, but not the HLA-B allele since its prediction confidence was low. If, however, the predicted HLA-C allele frequently occurs together with an HLA-B allele, we can use this relationship as a factor to boost the confidence of a prediction of HLA-B.

While treating each gene independently in the prediction has shown successful in prediction of common HLA alleles (the ones that are most frequent in the population), prediction of rare HLA alleles has been more challenging. Their representation in the population sample might not be sufficiently large for predictive models to generate a confident prediction. A common practice in HLA prediction is then to only predict, or make calls for the alleles that are predicted with high confidence, usually above a pre-specified confidence threshold. Knowing that there are strong patterns of linkage disequilibrium among HLA genes, and that certain alleles occur together more often than what is expected by random chance, provides an additional source of information which can be used in HLA prediction to improve prediction performance.

In Chapter 2 of this thesis we present a novel approach for prediction of HLA alleles from SNP data that uses the structural information that exists in the HLA region. We

[1] See Table 2.1 for definition of phase

demonstrate that accounting for the highly correlated nature of HLA genes during the prediction improves the prediction accuracy [53, 54]. The method accounts for high correlation among the HLA alleles by using HLA population haplotype frequencies [20]. These frequencies inform us about the co-occurrence of alleles inferred from a wide population and independent from the sample at hand. We use this information during the prediction as a weighted factor with a gene-specific classifier.

In addition, we describe two different ways that HLA gene dependency information can be incorporated into the prediction process and evaluate their impact on prediction accuracy. We show that adding this information improves the prediction performance on multiple real-world data sets and demonstrate that it is a valuable asset for HLA gene prediction.

1.2.2 Quantifying uncertainty in HLA data

In hematopoietic stem cell transplantation the selection of donors for a patient in need of a transplant is based primarily on HLA matching, and the lower the ambiguity of typing the easier it is to determine the probability of allele level match during the donor search [26]. In order to facilitate rapid identification of matched donors for HSCT several methods have been proposed to infer unknown phase and allele assignment [19, 35]. They typically employ statistical methods and the unique properties of the HLA region, such as its high linkage disequilibrium, in order to estimate haplotype frequencies and predict the most likely haplotype assignment for an individual with an HLA typing consisting of a set of ambiguous allele pairs. HLA typing with less ambiguity on average gives fewer high-probability phased high resolution haplotypes. We aim to measure per-gene ambiguity resulting from several HLA typing formats across four continental populations using an information theory-based measure, Shannon's entropy [71].

Some previous work has been done in measuring typing ambiguity – first a measure developed by Helmberg et al. [22], and more recently, the first application of Shannon's entropy to this problem by Cano [8]. Helmberg proposed a characterization of HLA typing kits using a frequency inferred typing (FIT) index. A FIT index describes the probability of correct allele pair assignment for an ambiguous typing result, and is calculated as the negative log of the probability of a wrong allele pair prediction. This probability is equal to the sum of products of all allele pairs that share the same typing

pattern as the selected pair. The limitation of the FIT index is that it does not take into account the distribution of allele frequencies beyond the most likely assignment. Both previous typing ambiguity studies used allele frequencies in their computations and, as we will later show, fail to demonstrate the advantage of linkage disequilibrium information contained in haplotype frequencies when it comes to reducing ambiguity and improving predictions of patient-donor matching.

In Chapter 3 of this thesis we describe an information-theory based measure to quantify ambiguity content in an HLA typing. We demonstrate that it can be objectively used to compare methods of HLA typing to each other in terms of the information they provide, in the context of each individual population. Our results show that intermediate-resolution single-pass sequence-based typing (SBT) contains the least ambiguity and, therefore, the most certainty in allele prediction across all populations. In addition, we demonstrate the benefit of using haplotype frequencies in entropy calculations versus allele frequencies. Neighboring HLA and non-HLA genes are highly correlated and major efforts have been directed at describing linkage disequilibrium (LD) across the region [11, 44, 80]. When certain alleles occur together generally due to linkage disequilibrium between them, some ambiguity can be inferred away using this linkage information, which is contained in haplotype frequencies. Our results show that using population haplotype frequencies greatly reduces the ambiguity present in HLA typing.

1.2.3 Prediction from uncertain HLA data

Allele and phase ambiguity found in most HLA data sets today presents a challenge to current SNP-based methods for prediction of HLA alleles, since they require non-ambiguous data. A way to obtain high-resolution HLA alleles from ambiguous data is to use one of the existing methods for resolving ambiguity from the ambiguous HLA data [36] and then use the generated high-resolution HLA alleles in learning the SNP-based prediction method. However, depending on the levels of ambiguity in the HLA data, the errors generated by these methods can be prohibitively high, up to 45% (this error includes the error in both, allele and phase assignment, and should be interpreted as the upper bound on the allele assignment error) [36]. These errors are then further propagated when the incorrectly imputed HLA alleles are used to train the SNP-based

prediction model.

In Chapter 4 we propose an approach to SNP-based prediction of HLA alleles that handles *allele* ambiguity in the HLA genotypes, and thus uses ambiguous HLA data as a learning source in the prediction [55]. Additionally, we evaluate how different levels of ambiguity in HLA data impact the prediction performance. Finally, we demonstrate that building the predictive model from ambiguous, rather than from statistically imputed high-resolution data, is a better approach to SNP-based prediction of HLA alleles.

1.3 Thesis Contributions to Computer Science

In this chapter we describe the contributions of this thesis to the field of computer science. The methods we developed and described in this thesis directly contribute to the areas of *correlated multi-label multi-class prediction* (Chapter 2) and *learning from uncertain data* (Chapters 3 and 4). We begin by describing existing work in these areas and conclude with limitation of existing methods to the problem at hand (the analysis and prediction of HLA genes).

1.3.1 Correlated multi-label multi-class prediction

Prediction of HLA alleles can be treated as a correlated multi-label multi-class prediction problem. Each HLA gene is a multi-class label. Labels are correlated because of the strong linkage disequilibrium in this region.

Multi-label classification is concerned with learning from a set of instances that are associated with more than one label, as opposed to traditional single-label classification problems, where instances are associated with a single class label. Multi-label classification problems are, nowadays, required by many applications, such as protein function annotation, image labeling, and document categorization. For a comprehensive review of techniques proposed in the area of multi-label classification and their applications refer to [77, 74].

The multi-label prediction methods most relevant to prediction of HLA alleles are those that account for the existence of dependence in the label set. The general agreement in the literature is that it is important to take into account label dependencies

during the classification process [3]. However, the problem of considering label dependency during prediction has an exponential nature. A number of approaches are, therefore, motivated by the computational demands of the correlated multi-label prediction problem. We next describe the most recent developments and applications of correlated multi-label prediction.

In protein function annotation, the goal is to predict functions of unannotated proteins in the context of molecular networks, where a protein can be annotated with a two or more functions which constitute the label set [28, 50, 76]. The algorithms in this category takes into account domain knowledge by combining the data (e.g. gene expression) and an existing structure related to the data, such as the Gene Ontology [2] or MIPS functional categories [3]. They then compute graph-topology related similarity between labels, that is, proteins in the graph, and use this similarity to propagate functional labels to unannotated proteins [28, 76]. The approach in [50] uses additional information about the genes (gene expression data) and combines the similarity between genes with respect to their expression and similarity between genes with respect to their relationship in the functional network. These approaches demonstrate that using the dependence between protein functions in the network can enable more accurate functional annotation.

Several approaches are proposed that model multi-label prediction as a binary relevance problem (BR) that incorporates label dependency by augmenting the feature space with predictions made on individual labels [60, 18, 3]. In the BR approach a multi-label classification approach is decomposed into multiple independent classification problems for each label separately. Each instance in the training set is labeled as positive if the given label is in the set of labels for that instance, and negative otherwise. The final label set for each instance in the data set is generated by aggregating results of classification over all individual classifiers. In [60] a chain of classifiers is formed to predict binary association for a single label on the feature space augmented by all previous BR predictions in the chain. Since the order of labels in the chain can change the prediction of the final label set, the authors address this issue by creating an ensemble of randomly ordered chains of classifiers and using predictions of each chain classifier

[2] http://www.geneontology.org/
[3] http://mips.helmholtz-muenchen.de/proj/ppi/

as votes. In [18], the authors create a two-stage classifier where they use SVM, as the first step, on the original feature set which returns predicted scores for each sample and each label independently. In the second step, the original feature step is augmented with these scores, which reflect the dependence between labels, and SVM is applied again. The method in [3] has a similar approach, but is not restricted to one underlying classification method (SVM in [18]), and also addresses possible over-fitting.

Several approaches predict correlated labels by augmenting the original label set with its power set. The most recent and promising approach in this category is described in [78]. They avoid the computational complexity of power set approaches for datasets with large numbers of labels by building an ensemble framework where each classifier is a single-label classifier for the prediction of a power set of a random small subset of labels. Lastly, generative approaches are proposed to learn joint distributions of features and labels. The most recent such approach is described in [15] where two graphical models are built to learn parameters associated with feature-label-label triplets, therefore, capturing the impact that an individual feature has on co-occurrence with a pair of labels. This approach proved efficient, but is restricted to pairs of labels only.

Challenges to Using Multi-label Prediction for HLA Prediction

There are several challenges, dictated by the nature of the data, to applying existing multi-label prediction methods to the HLA inference problem. As described earlier, most of the existing methods were developed in the areas of protein function annotation, text categorization and image annotation, where the existence of binary labels is assumed, e.g. a protein either has a specific function or not, a document either belongs to 'Sport' category or not, and so on. In addition, a data point can be assigned an arbitrary number of labels, e.g. an image can be classified as 'Sky' or it can be classified as 'Sky', 'Sea', and 'Summer', that is, there is no strict constraint on the number of labels a data point can be assigned. We next explain these and other issues in more detail.

- **Sparse label space**

  HLA prediction is a problem of predicting a small number of HLA labels (genes) from SNP data where each label has a large number of values (HLA alleles). In

a simple example, we are interested in predicting HLA-A, -B, and -DRB1 genes for a particular sample in which the number of known alleles for each gene is $N_A = 450$, $N_B = 300$, $N_{DRB1} = 200$, respectively. We cannot directly apply any of the existing methods that assume binary labels to this problem, as the labels are categorical. We can, however, convert categorical labels into binary, by creating N_A, N_B, N_{DRB1} binary labels instead of the three categorical ones. Each data point is assigned '1' for exactly three attributes corresponding to its three alleles, and '0' otherwise. Note that each data point is assigned exactly N labels, where N is the number of genes. This makes the label space extremely sparse, since HLA prediction is concerned with predicting several HLA genes, that is, predicting several binary labels out of hundreds or thousands of possibilities.

- **Non-arbitrary number of labels**

  Even though the conversion of categorical into binary labels takes care of the binary label assumption, we still run into problems with existing multi-label prediction algorithms due to their lack of constraint on the number of predicted labels. A data point can be assigned 1 or any arbitrary number of binary labels depending on the problem at hand and the parameter setup of the algorithm. For example, an image can be annotated with 'Sky', 'Sea', 'Summer', and 'Ocean' labels, but any subset of these labels would be a correct solution as well. For the prediction of N HLA alleles in a binary label set scenario, we are concerned with predicting exactly N binary labels for each subject. An additional constraint is that these N binary labels cannot be selected randomly from the label set, but each of them has to belong to a disjoint set of alleles describing each gene. In the example described above, the number of binary labels is $N_A + N_B + N_{DRB1}$, and we are concerned with predicting exactly *three* labels so that each one comes exclusively from N_A, N_B and N_{DRB1} alleles, respectively. Such a constraint is not required in applications like protein function annotation, text categorization and image annotation, and therefore not enforced in algorithms designed for these applications.

- **Lack of additional structure**

  In some multi-label classification problems, e.g. protein function annotation, the

classification is commonly assisted by an additional structure describing relationships among labels. For example, for two genes, one with known and one with unknown function, that have similar expression profiles, a protein interaction network is used to propagate the function along the edges and annotate the unknown gene. Taking such structures into account in multi-label prediction is important, because it can lead to improved predictive power and computational efficiency. Computing such additional structure on the binarized HLA label set would be very challenging due to its sparsity and high-dimensionality as mentioned above. In addition, since the relationships between alleles of the same gene are not useful, we only need to compute the relationships between labels belonging to different HLA genes, but not among them. This would result in an overall sparse structure with possibly many disconnected nodes (HLA alleles on one gene that are not correlated with any other alleles on other genes). Such a disconnected structure might not be as great an asset to the HLA inference problem as the traditional protein functional network is to the functional annotation of genes.

- **Need for SNP phasing**

  Lastly, humans are diploid organism, that is, we have two copies of each chromosome, one from the mother and one from the father. This has a direct implication in HLA prediction in that we need to predict alleles of both copies of HLA genes, which, due to the high polymorphism of this region, are more often different than not. SNP genotype data, which we start from, is reported in unphased form, that is, each person has one genotype (for definitions of phase and genotype, see Table 2.1). In order to impute two alleles for each HLA gene, we need the assistance of SNP data arranged on each chromosome separately, that is the SNP haplotype data. This is why all existing algorithms in HLA prediction either use phased SNP data as a starting point, or a phasing algorithm to infer SNP haplotypes before or during HLA prediction. The performance of SNP phasing will impact the follow-up analysis and, therefore, special attention needs to be taken at this step.

As seen above, HLA prediction cannot be safely treated as a traditional multi-label prediction problem due to its particular nature and many prediction constraints resulting from it. Solving HLA prediction with the help of multi-label prediction requires modification of existing algorithms or the design of new ones with HLA specific characteristics in mind. The method for correlated multi-label prediction of HLA genes that we describe in Chapter 2 directly address all of the above challenges. It can be easily applied to other problems of similar nature, such as SNP imputation or prediction of other genomic regions, or adapted to solve problems of simpler nature, such as any multi-label multi-class prediction that does not require genomic phasing.

1.3.2 Analysis and learning from uncertain data

Traditional data mining methods assume the data are complete and exact. However, in many applications data points are uncertain due to a number of factors including noise, measurement errors, aging data collection methods, etc. For example, in demographic data sets, the information about household income is often aggregated rather than known for every individual [1]. Uncertain data present a challenge to data mining techniques since most of them assume certainty and are not equipped to deal with inexact or noisy data.

Despite numerous classification algorithms for exact data, the work in the area of prediction from uncertain data remains limited. A modified support vector machine algorithm was proposed in [82] that handles input uncertainty. The modified algorithm adjusts the margin based on the uncertainty of data points which lie on the boundary. The authors showed that by modeling input uncertainty, they could obtain more accurate predictions when compared to the standard SVM classifier. A modified naive Bayes classification of uncertain data has been developed that extends the class conditional probability estimation in the Bayes model to handle probability distribution functions defined over uncertain data points [61]. The method was tested on several data sets that contain uncertainty and demonstrated that the proposed distribution-based method outperforms naive Bayes on a specific data set obtained by replacing each probability density function with its expected value. A decision tree based classification method has been developed for uncertain data by modifying partitioning measures, such as entropy

and information gain, to accommodate probabilistic nature of the attributes [58]. Similarly, a rule-based classification algorithm for uncertain data has been described that uses probabilistic information gain for generating rules [59]. For a comprehensive review of data mining techniques for storage and analysis of uncertain data refer to [2, 45].

We next describe the limitations of existing approaches to our application.

- All existing methods for learning from uncertain data are developed for one-label problems, and do not accommodate the multi-label applications, such as ours. Even if these methods were to be applied on our data one-label at a time, therefore reducing the problem to a one-label problem, there are still other limitations to using traditional classification approaches in our application, described in detail in the previous Section 1.3.1.

- Most of described existing methods assume numerically uncertain data points, or data points that can take on any value from a defined distribution or from a certain range of values [82, 61, 1]. However, uncertainty can also occur in categorical data. For example, a tumor is often difficult to accurately classify due to limitations in lab experiment precision. Therefore, doctors may often classify a tumor as benign or malignant with a certain probability [58]. In this thesis, we deal with categorical uncertainty in the form of uncertain HLA assignments.

In Chapters 3 and 4 we describe novel algorithms for the analysis and prediction of uncertain HLA data. Molecular typing of HLA data is often returned by the typing laboratories as an ambiguous set of possible alleles, all equally likely. In Chapter 3 we propose a novel entropy-based algorithm for quantifying the uncertainty in the data. In addition, we demonstrate that the correlation that exists among uncertain data, can eliminate some of the originally reported uncertainty. In Chapter 4 we propose a novel EM-based algorithm for prediction of uncertain categorical labels. The EM algorithm iterates through all samples and all possible resolutions of uncertain data points, to estimate the distribution of all possible outcomes. The framework can be easily adopted for prediction of uncertain categorical data in other applications.

1.4 Thesis Overview

The organization of this thesis is summarized in Table 1.1. We begin in Chapter 2 by describing correlated SNP-based prediction of HLA alleles. In Chapter 3, we describe a method to quantify ambiguity in HLA data, and demonstrate on simulated data that a method is well suited to compare common HLA typing methods. Chapter 4 describes a method that utilizes available ambiguous HLA data sets for SNP-based prediction of HLA alleles by incorporating the ambiguity into the training process. Finally, we conclude the thesis in Chapter 5 with discussion and proposed future work.

Table 1.1: Thesis Organization: Prediction and Analysis of HLA Data

	Exact data	**Uncertain data**
Prediction	Prediction of HLA data from SNPs (Ch. 2)	Prediction of HLA data from SNPs and ambiguous HLA (Ch. 4)
Uncertainty evaluation	Measuring ambiguity in HLA data using haplotype frequencies (Ch. 3)	

Chapter 2

Prediction of HLA Genes from SNP Data

2.1 Introduction

Obtaining HLA data for large disease-association studies may be of great value for learning about biological mechanisms behind many diseases. However, due to the high cost of HLA typing, obtaining that data for large studies may be impractical. In this chapter we present an approach for prediction of HLA genes from existing Single Nucleotide Polymorphisms (SNP) data with the goal of enriching available large genome-wide studies with this valuable new information. For terminology used in this and upcoming chapters, see Table 2.1.

2.2 Existing Work in HLA Imputation

Several methods have been developed and used to infer HLA alleles from SNP data [11, 32, 70, 43, 34]. We next divide them into several categories and briefly describe them.

Method that uses tag SNPs to infer HLA alleles

Initial attempts in HLA inference from SNP data proposed the use of tag SNPs to infer common HLA alleles [11]. This study observed that there is a strong linkage

Term	**Definition**	**Example**
Diploid	Having two copies of DNA or pairs of homologous chromosomes. Humans are diploid organisms.	
Phase	A physical ordering of markers (SNPs, alleles) along a chromosome.	
Haplotype	An arrangement of markers (SNPs, alleles) on a chromosome. When a physical arrangement of a set of alleles is known, we also say that the data is ***phased*** (or that the ***phase*** is known).	A T G C (sometimes written as $A \sim T \sim G \sim C$)
Genotype (un-phased genotype)	A set of un-phased markers for diploid organisms (no information on physical arrangement of alleles on chromosomes).	A/T T/T G/A C/T
Haplotype pair (phased genotype)	A physical ordering of markers (SNPs, alleles) along the two chromosomes in diploid organisms.	A T G C, T T A T (sometimes written as $A \sim T \sim G \sim C, T \sim T \sim A \sim T$)
kb = (kbp)	kilo base pairs $= 1,000bp$. Base pair(s), *bp*, is a unit commonly used to describe the length of a D/RNA molecule.	$50kbp$

Table 2.1: Terminology used in the thesis.

disequilibrium [13] between genetic markers and some HLA alleles and listed several such associations. Although the study demonstrated efficient prediction of some common HLA alleles using one or two SNP markers, this approach does not provide a solution to accurate HLA inference. HLA genes are highly polymorphic with many rare alleles, which are hard to predict using a small number of tag SNPs. Additionally, a large number of HLA alleles would require a large number of tag SNPs to be identified. Finally, the identification of tags in a small sample could cause over-fitting, since the selected SNPs might be sample-specific and not generalize well to future data.

Methods that use phased SNP haplotypes to infer HLA alleles

Another set of approaches aim to improve the performance of tag SNPs for HLA prediction by utilizing the joint distribution of SNPs in the MHC genomic region as reflected in SNP haplotypes [32, 70, 43]. In genetics, a haplotype, also known as a phased genotype, is a physical ordering of markers (SNPs, alleles) along a chromosome. Note that this arrangement needs to be inferred using statistical methods or pedigree information, since sequencing of the genome is performed simultaneously on both chromosomes, and an individual's genotype, and therefore, does not uniquely define their haplotype.

Leslie et al. [32] proposed HLA inference from SNP haplotypes. They assume prior availability of phased SNP and HLA data, either previously known or estimated. Their method is based on the assumption that a chromosome carrying an HLA allele is an imperfect mosaic of other chromosomes with the same allele. Given training data of SNP haplotypes and the corresponding phased HLA alleles, a hidden Markov model is used to compute the probability that a chromosome with a given SNP haplotype carries a particular HLA allele. This method requires a fine genetic map of input SNP data.

Similarly, approaches in [70, 43] rely on family trios to infer genotype phase more precisely. This information, however, is rarely available in large SNP data sets. Approaches in this category either start from SNP haplotype data or infer the phase from SNP genotypes and then impute HLA alleles. SNP haplotypes are not always available and need to be inferred using one of many genotype phasing methods [6]. The performance of phasing methods, therefore, impacts the follow-up analysis and HLA imputation.

Methods that use SNP genotypes to infer HLA alleles

In this set of approaches, HLA alleles are inferred from a set of selected SNPs in the MHC region [34]. Instead of using SNP haplotypes, the method described in [34] builds a predictive model for each HLA gene using unphased SNP data from a set of 630 individuals as the training set and tested it on a set of 630 unrelated individuals. This method uses a maximum likelihood framework based on all the HLA and SNP genotypes in the sample, identifies the most informative SNPs and computes predictive probabilities for all possible HLA allele pairs, given the SNP genotype data. The pair with the

maximum probability is predicted to be the HLA type for that sample if the probability exceeds a pre-specified confidence threshold (ct); otherwise, no prediction is made. Typically ct is set to zero, meaning all samples are predicted, that is, the call rate is 100%. When ct is increased, the accuracy generally increases. However, the call rate drops, in some cases to around 35%. Empirical evaluation of known issues in HLA inference is described in [84] and demonstrated with the methods in [34] on several datasets of different ancestries and different typing methodologies. Lastly, the application of the proposed inference method on several other polymorphic genes is shown in [83].

2.3 Limitations of existing work

All existing approaches make predictions for HLA alleles for each gene independently. They consider each HLA gene as an independent label in a multi-label prediction scenario, and build predictive models with respect to each label separately. Since many HLA alleles are known to be in linkage disequilibrium [13], we can take advantage of this label dependency to assist the HLA inference and potentially improve its accuracy. This is especially relevant since many of the existing algorithms use a confidence threshold (ct) to decide whether they want to make a prediction or not. If ct is set to zero, all samples are predicted and the accuracy of prediction is typically lower. If ct is set to a number greater than zero, the accuracy of prediction increases, but not all samples are assigned HLA alleles. Using the relationship between HLA alleles can boost the confidence of the prediction and increase the call rate on the entire data sample. For example, we might predict an individual's HLA-A allele with high confidence, but not the HLA-B allele since its prediction confidence was low. If, however, the predicted HLA-A allele frequently occurs together with an HLA-B allele, we can use this relationship as a factor to boost the confidence of a prediction of HLA-B. The flowchart of the HLA imputation using HLA dependence in prediction is shown in Figure 2.1.

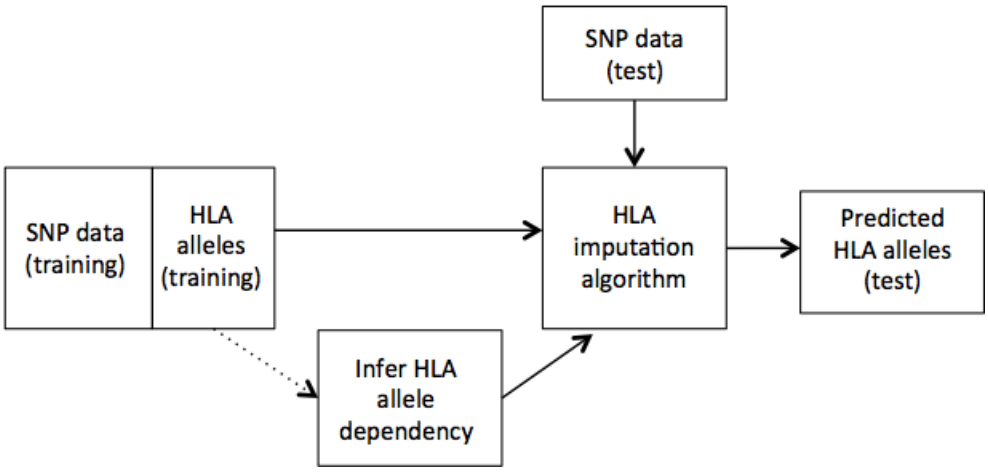

Figure 2.1: Flowchart of correlated HLA imputation from SNP data. The dashed line indicates that label dependencies can be inferred from the training sample or from a larger population, as is the case in our proposed work.

2.4 Proposed Approach

Prediction of HLA alleles from SNP data has been more or less successfully applied. Existing methods treat each gene independently and train a predictive model on each gene to predict an allele on that gene. While this approach has been successful in prediction of common HLA alleles (the ones that are most frequent in the population), prediction of rare HLA alleles has been more challenging. Their representation in the population sample might not be sufficiently large for predictive models to generate a confident prediction. A common practice in HLA imputation is then to only predict, or make calls for the alleles that are predicted with high confidence, usually above a pre-specified confidence threshold. Knowing that there are strong patterns of linkage disequilibrium among HLA genes, and that certain alleles occur together more often than what is expected by random chance, provides an additional source of information which can be used in HLA imputation to improve prediction performance.

We use the population haplotype frequency to inform us about the co-occurrence of alleles inferred from a wide population and independent from the sample at hand. We use this information during the prediction as a weighted factor with a gene-specific

classifier.

2.4.1 Top k predictions

We first select a set of possible HLA allele predictions by using SNP genotype data. We do this by running genotype data through an EM classifier and selecting the top k predictions for each gene independently. To obtain the best k guesses we use SNPs around each HLA gene separately, that is, to classify HLA-A alleles, we use SNPs around this region only, and so on. This step is shown in Figure 2.2. We discuss the details of this step later in this section.

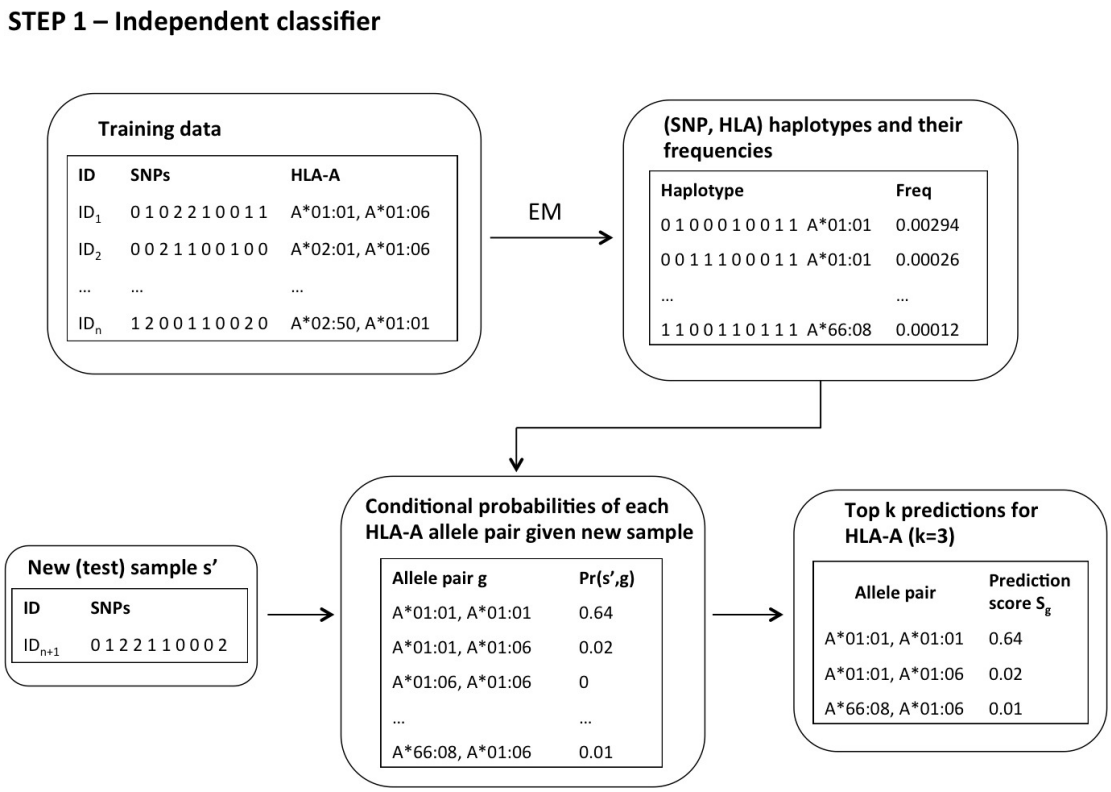

Figure 2.2: To obtain top K predictions for each HLA gene, we compute the probability of the test sample having every possible pair of alleles for a given gene and keep the K predictions with the highest probabilities.

For the prediction of each gene, we select a subset of SNPs within and around that gene. For this study we use a region of $\pm 50kb$ around each gene (for definition of kb refer to Table 2.1). We then train a classifier using these SNPs only to select the top k predictions for each gene. Note that genotype data compresses the sequence of two chromosomes into one vector, so that each of the k predictions will be a pair of HLA alleles, one for each chromosome, as shown in Figure 2.2. For example, the top $k = 3$ predictions for gene HLA-A in Figure 2.2 are

$A*01:01, A*01:01$
$A*01:01, A*01:06$
$A*66:08, A*01:06$

In the next step we enforce additional structure in the allele (label) space to refine the search and predict HLA alleles.

2.4.2 Population HLA haplotype frequencies

HLA genes are correlated because of the strong linkage disequilibrium in this region. We can utilize the structure that exists between the genes as an additional source of information in order to improve the prediction.

Given a set of likely predictions of a new sample with unknown HLA alleles found in the previous step, we use population haplotype frequencies [37] to narrow down the selection of HLA alleles and make final predictions. The haplotype frequencies are estimated at the level of the entire US population and inform us of how likely sets of alleles are to co-occur in each separate population [37]. We utilize the structure that exists between the labels as an additional source of information in order to improve the prediction.

Given a set of HLA haplotypes $h_j = (A \sim C \sim B \sim DRB1 \sim DQB1)$ and their population specific frequencies f_j, the frequency of any subset of genes (in this example genes C and B) can be calculated as

$$f(C^J \sim B^K) = \sum_j f_j(A \sim C^J \sim B^K \sim DRB1 \sim DQB1) \tag{2.4.1}$$

Now, given any number of HLA genes, a table similar to Table 2.2 can be constructed

to describe the dependency between those genes. This will become relevant when we use local haplotype information to inform the final predictions. Note that we here use a common notation of haplotypes which is a sequence of genes separated by the $\sim$ sign.

Table 2.2: HLA haplotype frequency table containing haplotypes and the frequencies with which they occur in a given population.

A	C	B	DRB1	DQB1	Frequency
01:02	07:01	07:02	07:01	02:01	6.79×10^{-7}
01:02	07:01	07:02	15:01	06:02	1.032×10^{-5}
01:02	07:01	08:01	11:02	03:01	1.031×10^{-5}
...	...	...	...	...	...

2.4.3 Combining top k predictions and haplotype frequencies

Once we obtained the top predictions of allele pairs and constructed the HLA haplotype frequency table, we combine the two to make a prediction. This step is shown in Figure 2.3.

Given a set of selected allele pairs for each gene, we generate all possible pairs of HLA haplotypes. Let's call this set H. For example, given the top $k = 3$ predictions for genes HLA-A, -B, and -DRB1 shown in Figure 2.3, some of the possible HLA haplotype pairs combining those predictions are the following:

$$(h_1, h_2) = \begin{cases} A*01:01 \sim B*07:01 \sim DRB1*01:01, A*01:01 \sim B*08:02 \sim DRB1*03:01 \\ A*01:01 \sim B*08:01 \sim DRB1*01:01, A*01:01 \sim B*13:02 \sim DRB1*03:01 \\ A*01:01 \sim B*07:01 \sim DRB1*01:01, A*01:01 \sim B*15:01 \sim DRB1*03:01 \\ A*01:01 \sim B*07:01 \sim DRB1*01:01, A*01:06 \sim B*08:02 \sim DRB1*03:01 \\ \qquad\qquad\qquad\qquad\qquad\qquad ... \\ A*66:08 \sim B*07:01 \sim DRB1*01:01, A*01:06 \sim B*15:01 \sim DRB1*10:01 \end{cases}$$

Each of the k predictions comes with a prediction score, let us denote it as S_A. This score could be prediction confidence, similarity with the test sample, or some other statistic returned by the classifier. These scores are assigned to individual alleles in the following manner: if prediction score of (A_1, A_2) is 0.4, both alleles A_1 and A_2 are individually assigned 0.4.

Given a haplotype frequency table constructed from the population haplotype frequencies, we retrieve the frequency of each HLA haplotype from H and assign a score

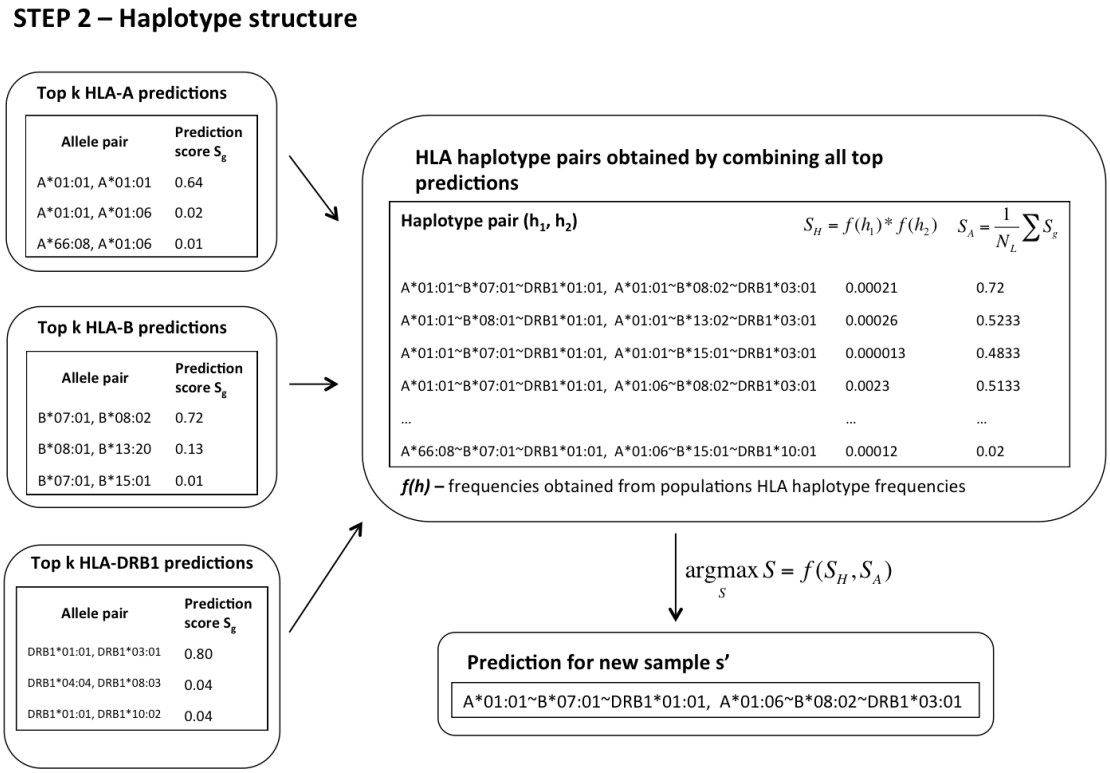

Figure 2.3: To incorporate haplotype structure into the prediction, we combine top prediction into all possible pairs of haplotypes and use haplotype population frequencies to refine the top prediction.

to each pair of haplotypes by multiplying their respective frequencies. In the above example, if haplotype $h_1 = A*01:01 \sim B*07:01 \sim DRB1*01:01$ has a frequency f_1 and haplotype $h_2 = A*01:01 \sim B*08:02 \sim DRB1*03:01$ has a frequency f_2, then the frequency of the haplotype pair is equal to $f_1 * f_2$. This frequency is labeled S_H. Here we assume that the two haplotypes occur independently in an individual. This assumption is known as a Hardy-Weinberg equilibrium or the assumption of random mating in a population.

To select the most likely pair of haplotypes for a test sample with unknown HLA alleles, we combine S_A and S_H into a final haplotype score S and make a prediction by selecting the highest ranking pair of haplotypes. For now, we denote the prediction

score as a function of S_A and S_H as follows

$$S = f(S_A, S_H) \tag{2.4.2}$$

The method is shown in Algorithm 1.

Algorithm 1 HLA Imputation

for each new sample s **do**
  Construct gene-specific classifiers
  Find k top predictions for s for each gene
  Assign prediction likelihoods as allele scores of top k allele pairs
  Using top predictions, generate all possible pairs of HLA haplotypes
  Retrieve haplotype frequency for candidate haplotypes
  Compute a score that combines allele and haplotype scores
  Determine a pair of haplotypes such that the overall score is maximized
end for

2.4.4 Contributions of the proposed approach

Here we list the contributions of our approach.

1. Instead of treating HLA genes independently and predicting alleles separately for each gene, we propose an approach that takes advantage of high linkage disequilibrium in the region. Additional information brought in by haplotype frequencies describing allele co-occurrence on the population level improves the prediction. Additionally, this structure brings into the prediction information about the general population that cannot be deduced from the sample at hand. It makes our algorithm more generalizable to samples of different sizes or ethnic backgrounds.

2. Our approach makes predictions for all test samples regardless of the confidence thresholds used for the base classifier.

3. An additional advantage of our approach is the prediction of all HLA haplotypes, rather than separate alleles. Our algorithm can be easily extended to return a ranked list of HLA haplotypes that a new sample is likely to have. This is valuable in the context of stem cell transplantation where donor search is performed based on the highest ranking imputed haplotype pairs.

2.5 Implementation of the approach

As described in the previous section, the first step in the prediction of HLA alleles of a new sample is predicting the top k allele pairs by running a classifier on SNP genotype data. We implemented this step by using an *Expectation-Maximization (EM)* classifier which has been extensively used in estimation of haplotype frequencies [37, 29]. We use EM to model the joint distribution of SNP and HLA alleles by estimating the frequency of the their joint haplotypes, and then use Bayes' theorem to compute the probability of a pair of HLA alleles given a new sample with SNP genotype data. We describe this methodology in more detail in the following section.

2.5.1 Selecting the top k allele pairs

Building the model using training data

The EM algorithm has been described in detail in [29], and we summarize it here briefly. Given HLA and SNP genotypes for all samples in training data set, EM estimates the frequencies of all haplotypes that could be obtained from observed genotypes, using the following two steps:

- *E-step (Expectation)* - Current haplotype frequencies are used to calculate for each subject conditional probabilities of each possible pair of haplotypes. These are then used to update current frequencies of each possible pair of haplotypes.
- *M-step (Maximization)* - The updated haplotype pair frequencies from previous step are used to update haplotype frequencies. In the initial step, haplotype frequencies are set to $1/n_h$ where n_h is the number of unique haplotypes in the data set.

These steps are also shown in Figure 2.4. In the first step EM uses the entire training data set and expands the training genotypes into all possible pairs of haplotypes (essentially phasing the data or assigning the observed genotypes to the two chromosomes). An example of this expansion for one sample is shown in Figure 2.5. EM expands the entire training data set and collects all unique (SNP, HLA) haplotypes from the set of expanded haplotype pairs. It assigns initial frequencies of all haplotypes to be equal

($1/n_h$, where n_h is the number of unique haplotypes). EM then iterates through the *E-step* and *M-step*, using the equations shown in the Figure 2.4, until the estimated haplotype frequencies do not change beyond a certain threshold or a pre-specified number of iterations is exceeded.

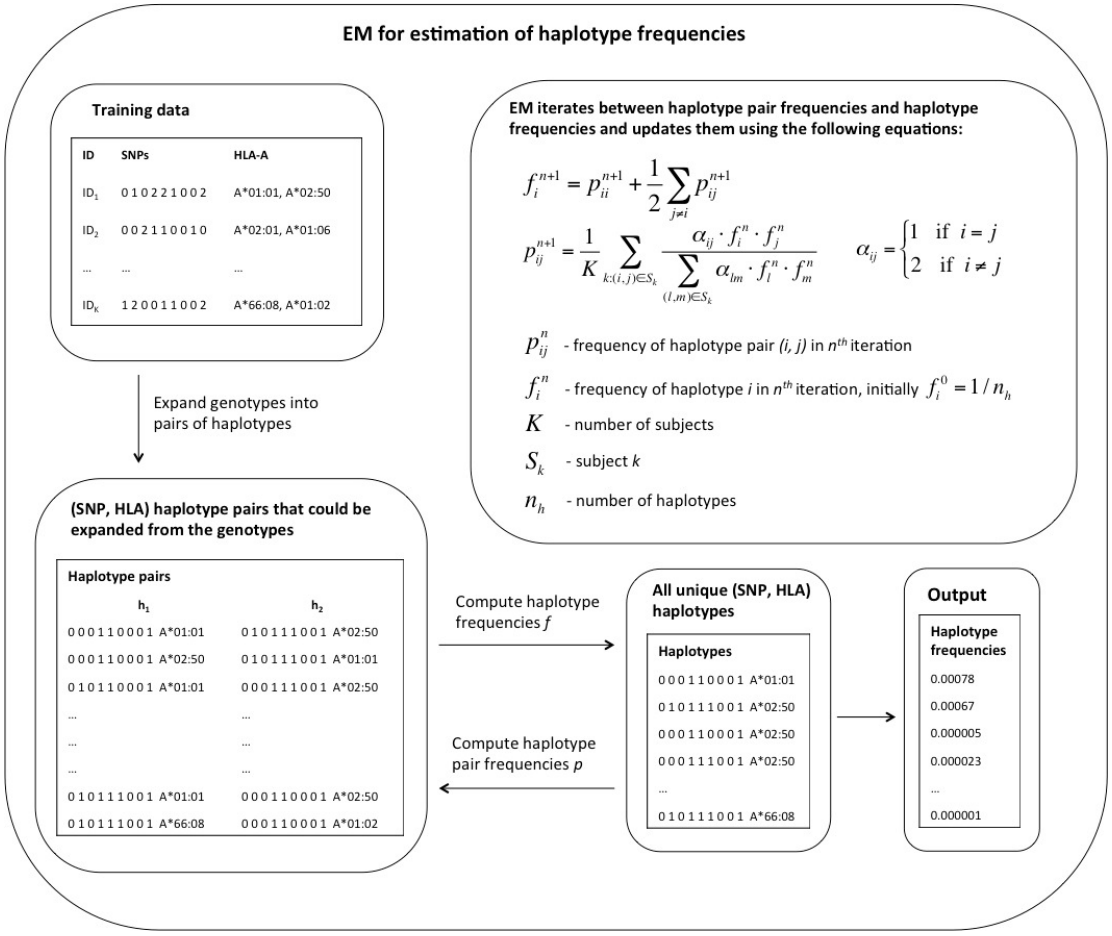

Figure 2.4: Expectation-Maximization (EM) algorithm for estimation of haplotype frequencies.

The outcome is a set of joint (SNP, HLA) haplotypes $h = (g_a, s_{a_1}, s_{a_2}, ..., s_{a_m})$ that could be obtained from the observed genotype data in the training sample and their estimated frequencies $f(g_a, s_{a_1}, s_{a_2}, ..., s_{a_m})$ where g_a denotes a gene allele, s_a denotes a

SNP allele and m is the number of SNPs.

Making a prediction for a test sample

Let $g_i = (g_{i1}, g_{i2})$ denote an HLA genotype of each subject i and $s_i = (s_{ij}, \forall j \in [1, m])$ represent a SNP genotype and let g_a and s_a denote a gene and a SNP allele, respectively. Given estimated haplotype frequencies $f(g_a, s_{a_1}, s_{a_2}, ..., s_{a_m})$, and a new sample with a SNP genotype $s' = (s'_j, \forall j \in [1, m])$, the probability that this sample has any two HLA alleles $g = (g_1, g_2)$ can be calculated as follows

$$P(g|s') \propto P(g, s') \tag{2.5.1}$$

where $P(g, s') = P(g_1, g_2, s')$ is a joint probability of HLA and SNP genotypes, and can be obtained by summing the frequencies over the set H of all (HLA, SNP) haplotype pairs that could be expanded from that genotype, as follows:

$$P(g, s') = P(g_1, g_2, s_1, s_2, ..., s_m) = \sum_H f(g_1, s_{a_{11}}, s_{a_{21}}, ..., s_{a_{m1}}) * f(g_2, s_{a_{12}}, s_{a_{22}}, ..., s_{a_{m2}}). \tag{2.5.2}$$

An example of expanding a new SNP genotype s' and an allele pair $g = (A*01 : 01, A*02 : 50)$ into all possible pairs of haplotypes and computing their joint probability is shown in Figure 2.5.

The predicted HLA allele pair for a new SNP genotype s' is that for which the probability $P(g|s')$ is maximal:

$$(g'_1, g'_2) = \arg \max_{(g_1, g_2)} P((g_1, g_2)|s'). \tag{2.5.3}$$

The prediction score is assigned to each of the two selected alleles, in order to obtain an allele score. We use the allele scores later along with population haplotype frequencies to make the final predictions.

$$S_{g'_1} = S_{g'_2} = P((g_1, g_2)|s') \tag{2.5.4}$$

Our confidence in this prediction is dictated by the maximal probability. In order to keep strong predictions and reduce the complexity of the problem later on, we do the

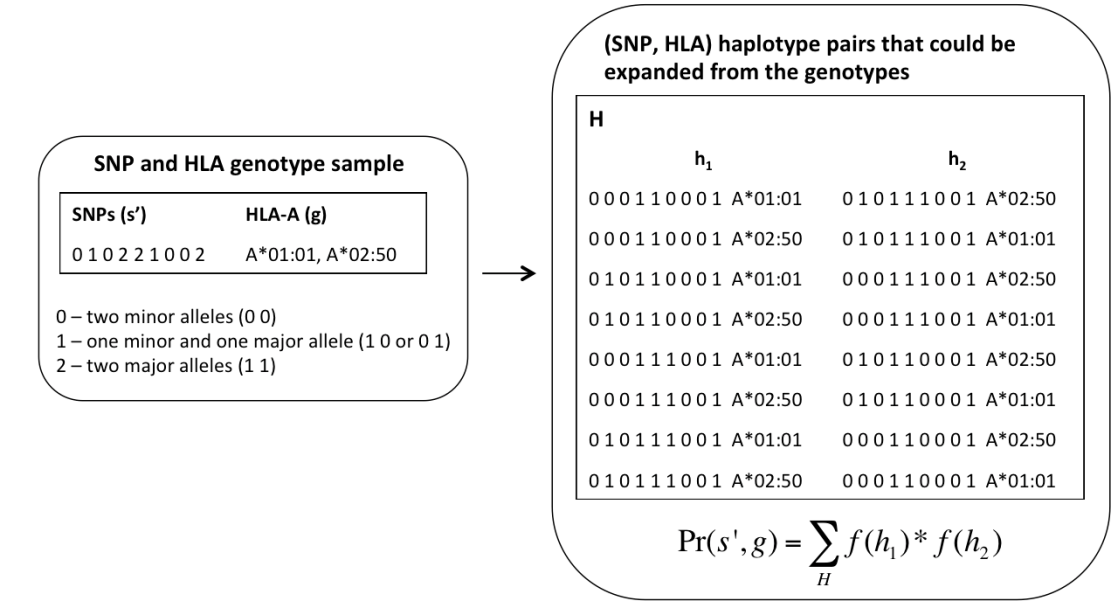

Figure 2.5: An example of expansion of genotypes into pairs of haplotypes.

following:

- If highest prediction probability exceeds pre-set threshold ct, we keep top *one* pair of alleles.
- If highest prediction probability does not exceed ct, we keep top k allele pairs.

We do this in an HLA gene-specific manner, that is, using SNPs around each HLA gene separately, to obtain top predictions for each sample in the testing data and for all loci $L \in \{HLA-A, -C, -B, -DRB1, -DQB1\}$.

We then move on to incorporating HLA haplotype frequencies into our final prediction.

2.5.2 Prediction score

Once we have selected allele pairs for each gene, we list all possible haplotype pairs $h = (h_1, h_2)$ resulting from these alleles. Each of these haplotype pairs is assigned an *allele score* S_A and an *haplotype score* S_H as follows:

$$S_A = \frac{1}{N_L} \sum_g S_g \tag{2.5.5}$$

$$S_H = f(h_1) * f(h_2) \tag{2.5.6}$$

where the summation is over all alleles g in haplotype h, and $f(h)$ is the haplotype frequency of h. Each haplotype pair h is then assigned a weighted score

$$S = (1 - \delta)S_A + \delta * S_H \tag{2.5.7}$$

where $\delta \in \{0, 1\}$.

The predicted pair of haplotypes for a new sample is the one with the highest score (Figure 2.3). For small δ, predictions are based more on allele score S_A and less on haplotype score S_H, that is the population haplotype frequencies, and vice versa.

2.5.3 Results

Data sets: We used HLA and SNP data from the International HapMap project[4]. This data set contains 90 individuals of African ancestry (YRI), 180 individuals of European ancestry (CEU), 89 individuals of Asian ancestry (ASI) - 44 Japanese (JPT) and 45 Chinese (HCB). These individuals are typed at 6,300 SNPs in the MHC region and 6 HLA loci (HLA-A, -C, -B, -DRB1, -DQA1, -DQB1).

The haplotype frequencies we used in this study were published in [37]. They contain 7,987 haplotypes and their relative frequencies in African, European, Asian and Hispanic populations.

Experimental setup: For each data set, we randomly selected 80% of samples to use as a training set and 20% of samples to use as a test set. We then used the proposed approach to predict the HLA alleles for all test samples and computed the accuracy of

[4] http://hapmap.ncbi.nlm.nih.gov/

the prediction (for each gene separately) by comparing the obtained alleles on the two predicted haplotypes with the true alleles for that sample. More precisely, we computed accuracy as follows:

$$accuracy_L = \frac{\sum |P_L \cap T_L|}{2 * N_T} \tag{2.5.8}$$

where the summation is over all test samples. P_L and T_L are predicted and true alleles at that gene, respectively, and N_T is the number of test samples. We additionally computed the overall accuracy of prediction across all HLA loci and across all test subjects. This accuracy is equal to the percentage of correctly predicted alleles across all HLA loci:

$$accuracy_{overall} = \frac{1}{N_L} \sum_L accuracy_L \tag{2.5.9}$$

where P and T are predicted and true alleles for all HLA loci.

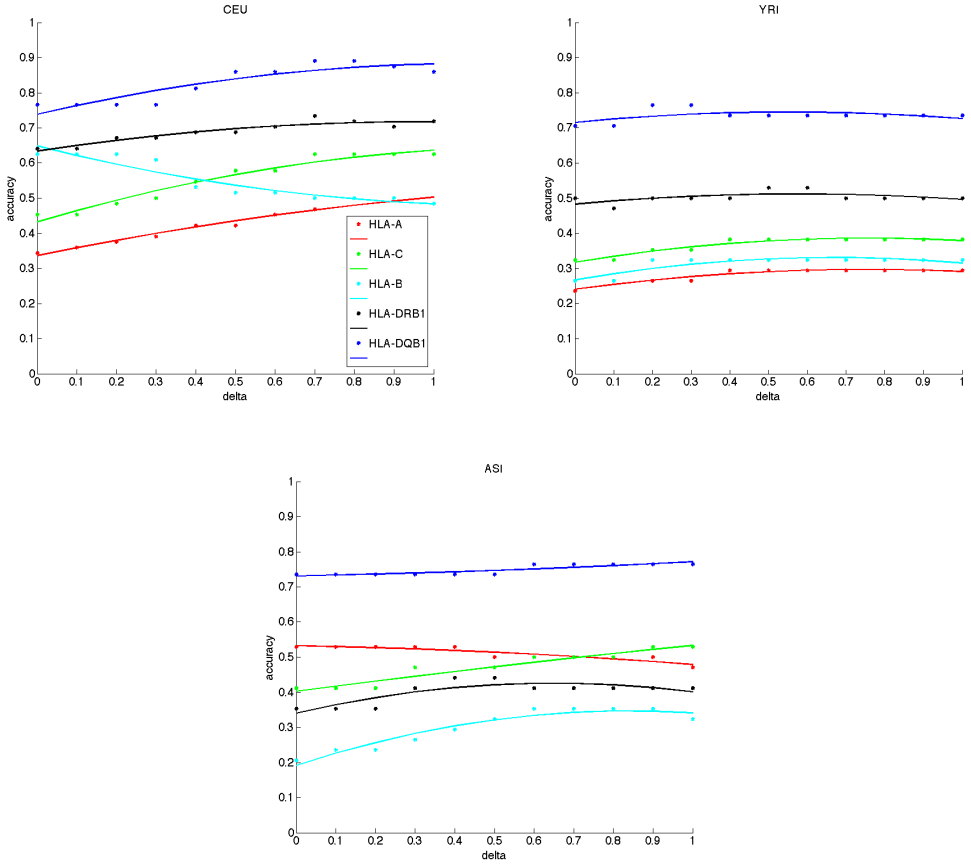

Figure 2.6: Prediction accuracy versus parameter *delta*. This figure illustrates the need to use HLA allele dependence to assist the prediction of HLA alleles. Figures 2.6(a), 2.6(b), 2.6(c) show gene-specific accuracy and correspond to the three study populations: CEU (European), YRI (African), and ASI (Asian). Different population frequencies are used for each population group. The color legend for 2.6(a), 2.6(b), 2.6(c) is shown in 2.6(a).

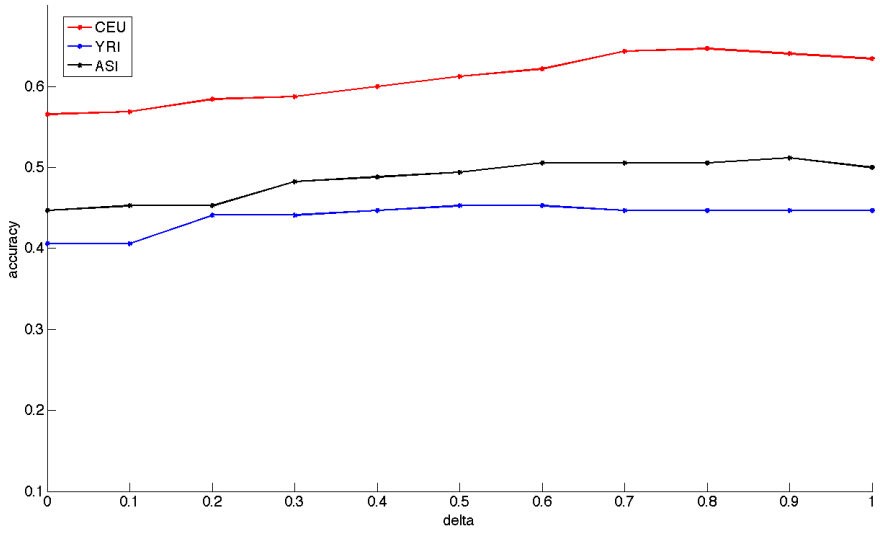

Figure 2.7: Overall accuracy across all HLA genes for each population.

We varied the threshold δ to demonstrate the need for using haplotype frequency in allele prediction. Figure 2.6 shows gene-specific prediction accuracy (Figures 2.6(a), 2.6(b), 2.6(c)) and the overall accuracy (Figure 2.5.3) for each study population using the proposed method. We used $k = 3$ top predictions in the first step, and varied δ. The figures demonstrate that as we increase δ, the accuracy generally increases. This means that we positively assist the prediction if we add the haplotype frequencies. For $\delta = 0$ the prediction is based solely on allele score which is computed from genotype data independently of haplotypes and allelic relationships.

It is evident from Figure 2.6 that improvement in accuracy varies for different loci and population groups. Gene HLA-B is the most polymorphic gene in the HLA system and therefore it's prediction tends to be the most challenging. We see that the accuracy for HLA-B is increasing with δ in the ASI population (Figure 2.6(c)), but not in the CEU (Figure 2.6(a)). In the next section we investigate how linkage or correlation between subsets of genes (local linkage) rather than all genes (global linkage) impact the prediction performance.

2.6 Evaluation of Label Dependency for the Prediction of HLA Genes

In the previous sections we demonstrated that accounting for the highly correlated nature of HLA genes during the prediction improves the prediction accuracy [53]. In this section we propose an improvement of the original method that more systematically deals with the selection of top predictions in the first step of the algorithm. Additionally, we propose two different ways that HLA gene dependency information can be incorporated into the prediction process and evaluate their impact on prediction accuracy. Next two sections describe these novelties.

2.6.1 Selecting top allele predictions

The improvement to the original method is introduced in the first step of the method that deals with selection of top predictions. In our previous work, we used a fixed k to select top k prediction. We now propose a more systematic approach that uses a cumulative likelihood of predictions to select top most likely ones.

As the first step in the algorithm, we select a set of possible HLA allele predictions by using SNP genotype data. We do this by running genotype data through a classifier and selecting the top most likely predictions. This was done for each gene separately to obtain top predictions for each sample in the testing data and for all genes ($HLA-A, -C, -B, -DRB1, -DQB1$), as shown in Figure 2.8. Note that the predictions for each gene will now have a different set of prediction candidates.
In order to keep the top predictions we do the following:

- Sort the predictions in descending order of their prediction likelihood.
- Move down the sorted list until the cumulative prediction likelihood exceeds a prespecified threshold t.
- Keep the visited predictions including the one visited last.

For example, given the output of the classifier for gene HLA-A shown in Table 2.3, and a thresholds $t = 0.95$, we keep the first *three* predictions, since their cumulative likelihood equals 0.96 and as such exceeds the threshold t.

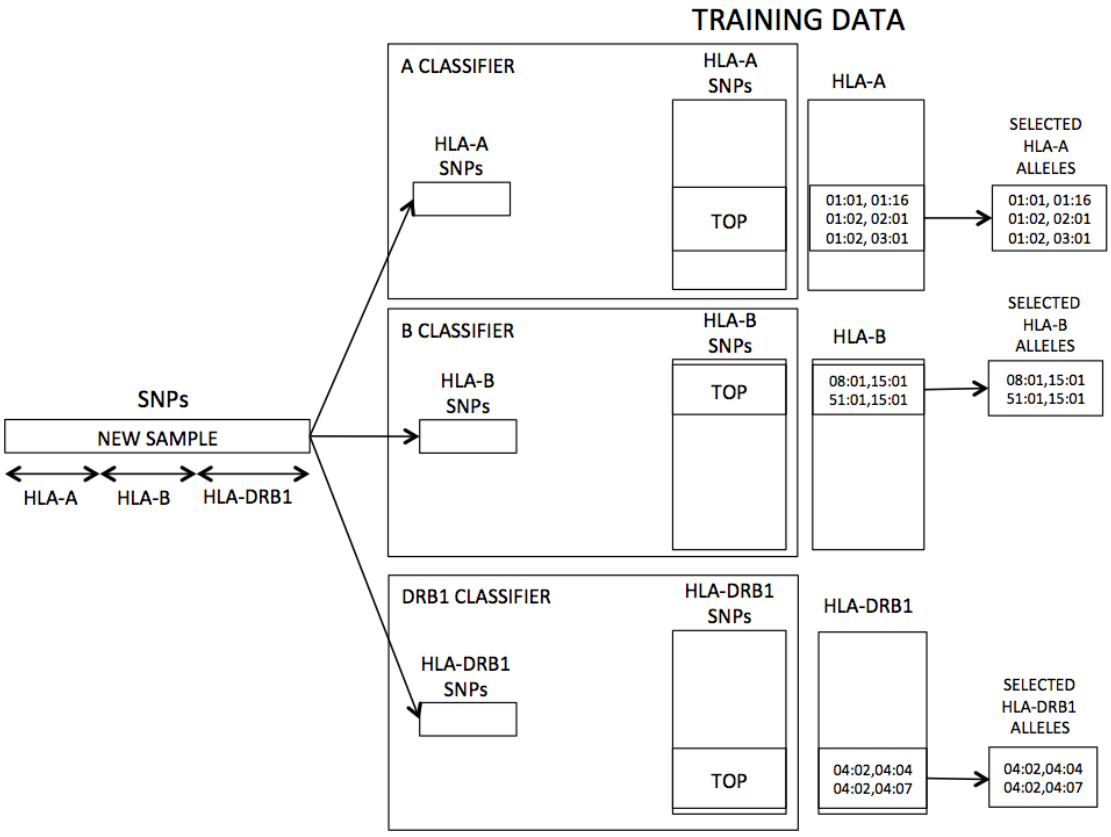

Figure 2.8: To obtain top predictions for each HLA gene, we compute the probability of the test sample carrying any pair of alleles for a given gene and keep the ones with the highest probabilities.

2.6.2 Combining allele and haplotype information into a prediction

Global haplotype structure

Once we obtained the top predictions of allele pairs and constructed the HLA haplotype frequency table, we combine the two to make a prediction. In the *global haplotype structure* approach, we combine predictions of all HLA genes into haplotypes and use their corresponding frequencies in the same manner as in Section 2.4.3. We briefly summarize it here.

Each of the top predicted HLA allele pairs obtained by EM comes with a prediction score, denoted as S_g. These scores are assigned to individual alleles on resulting

Table 2.3: An example output of the base (EM) classifier for HLA-A gene.

prediction	likelihood
A*01:01, A*01:16	0.8
A*01:02, A*02:01	0.1
A*01:02, A*02:01	0.06
A*01:02, A*24:02	0.04

candidate haplotypes. Each haplotype pair $h = (h_1, h_2)$ from a set of haplotypes H generated from top predictions is assigned an *allele score* S_A and *haplotype score* S_H as follows:

$$S_A = \frac{1}{N} \sum_g S_g \tag{2.6.1}$$

$$S_H = f(h_1) * f(h_2) \tag{2.6.2}$$

where N is the number of genes in the haplotype and the summation is over all alleles g in haplotypes, and $f(h)$ is haplotype frequency of h. To select the most likely pair of haplotypes for a test sample with unknown HLA alleles, we combine S_A and S_H into a weighted score S as follows:

$$S = f(S_A, S_H) = (1 - \delta) * S_A + \delta * S_H \tag{2.6.3}$$

where $\delta \in [0, 1]$.

The predicted pair of haplotypes for a new sample is the one with the highest score S. For small δ, predictions are based more heavily on allele score S_A and less on haplotype score S_H, and vice versa. We term this approach ***global haplotype structure*** to emphasize that the prediction is based on the global all-gene haplotypes, that is, using the dependency information over the entire set of labels (genes).

Local haplotype structure

Genes that are closer to each other on the chromosome exhibit stronger linkage disequilibrium than genes that are further apart. In the context of HLA, genes C and B are more strongly linked with each other than they are with gene A. The same is true for genes DRB1 and DQB1. This is because of their proximity on the chromosome as can be seen in Figure 1.1.

Here we investigate how the local linkage, or the ***local haplotype structure***, as we refer to it here, between genes C and B and genes DRB1 and DQB1 affects the prediction of C and B, and DRB1 and DQB1 genes, respectively. For the simplicity of notation, we refer to genes DRB1 and DQB1 as DR and DQ in the equations to follow. We use local haplotypes ($C \sim B$ and $DRB1 \sim DQB1$) instead of global ($A \sim C \sim B \sim DRB1 \sim DQB1$) haplotypes, as the second step in prediction. To obtain local haplotype frequencies, we use Equation 2.4.1.

The selection of top predictions at each HLA gene along with their alleles score S_g is the same as in the previous section. Once we have selected allele pairs for each gene, we list all possible local HLA haplotype pairs

$$h_{(C\sim B)} = (h_{(C\sim B)_1}, h_{(C\sim B)_2}) \tag{2.6.4}$$

and

$$h_{(DR\sim DQ)} = (h_{(DR\sim DQ)_1}, h_{(DR\sim DQ)_2}) \tag{2.6.5}$$

resulting from these alleles. Each of these haplotype pairs are assigned haplotype scores based on their frequencies, as follows:

$$S_{H(C\sim B)} = f(h_{(C\sim B)_1}) * f(h_{(C\sim B)_2}) \tag{2.6.6}$$

$$S_{H(DR\sim DQ)} = f(h_{(DR\sim DQ)_1}) * f(h_{(DR\sim DQ)_2}) \tag{2.6.7}$$

Similarly, each obtained HLA haplotype pair is assigned an *allele score* S_A based on the individual allele prediction probabilities S_g, as follows:

$$S_{A(C\sim B)} = \frac{1}{2} \sum_{g\subseteq\{C,B\}} S_g \tag{2.6.8}$$

$$S_{A(DR\sim DQ)} = \frac{1}{2} \sum_{g\subseteq\{DR,DQ\}} S_g. \tag{2.6.9}$$

where the summation is over all alleles g that occur in predicted haplotypes. $S_{A(x)}$ is used to denote overall *allele score* at the gene $x \in \{A, C, B, DRB1, DQB1\}$. Note that the allele score for gene HLA-A is based solely on allele prediction likelihood S_g, that is

$$S_{A(HLA-A)} = S_{g(HLA-A)} \tag{2.6.10}$$

We use full gene name HLA-A in the above equation to avoid confusion with allele score notation represented with A.

To select the most likely pair of haplotypes for a test sample with unknown HLA alleles, we combine allele and haplotype scores into a weighted score S as follows:

$$S_{HLA-A} = S_{A(HLA-A)} \tag{2.6.11}$$

$$S_{C\sim B} = f(S_{A(C\sim B)}, S_{H(C\sim B)}) = (1-\delta) * S_{A(C\sim B)} + \delta * S_{H(C\sim B)} \tag{2.6.12}$$

$$S_{DR\sim DQ} = f(S_{A(DR\sim DQ)}, S_{H(DR\sim DQ)}) = (1-\delta) * S_{A(DR\sim DQ)} + \delta * S_{H(DR\sim DQ)} \tag{2.6.13}$$

where $\delta \in [0, 1]$. The prediction of HLA-A gene is based solely on the allele score, while the prediction of HLA-C and HLA-B is based on both $C \sim B$ allele and haplotype scores, and the prediction of HLA-DRB1 and HLA-DQB1 is based on both $DRB1 \sim DQB1$ allele and haplotype scores.

2.6.3 Results

We perform a thorough evaluation of how global and local haplotype frequencies assist the prediction of HLA genes, using different δ thresholds and cross-validation.

2.6.4 Data sets

We used HLA and SNP data from the International HapMap project[5]. This data set contains 180 individuals of European ancestry (CEU) typed at $6,300$ SNPs in the MHC region and 6 HLA genes (HLA-A, -C, -B, -DRB1, -DQA1, -DQB1).

We also used SNP and HLA genotype data from the British 1958 Birth Cohort (BC)[6] downloaded from the European Genotype Archive[7]. Genotyping of the BC data was carried out using Affymetrix 6.0 platform [7]. High resolution typing of five HLA genes (HLA-A, -C, -B, -DRB1, -DQB1) is obtained using Sequence Specific Oligonucleotide

[5] hapmap.ncbi.nlm.nih.gov
[6] www.b58cgene.sgul.ac.uk
[7] www.ebi.ac.uk/ega

and Sequence Specific Primer methodologies[8]. After preprocessing, this data set consists of 1186 samples of European origin with all 5 HLA genes and 1795 SNPs across the MHC region.

The haplotype frequencies we used in this study were published in [37]. They contain 7,987 5-gene haplotypes and their relative frequencies in African, European, Asian and Hispanic populations. In this study we used European frequencies.

2.6.5 Experimental evaluation

For each data set, we randomly selected 80% of samples to use as a training set and 20% of samples to use as a test set. We then used the proposed approach to predict the HLA alleles for all test samples and computed the accuracy of the prediction (for each gene separately) by comparing the obtained alleles on the two predicted haplotypes with the true alleles. More precisely, we computed accuracy as follows:

$$accuracy = \frac{\sum_{i=1}^{N} |P_i \cap T_i|}{2 * N} \tag{2.6.14}$$

where the summation is over all test samples. P_i and T_i are predicted and true alleles of sample i, respectively, and N is the number of test samples. All experiments are conducted at the intermediate 2-digit resolution of HLA alleles, and repeated for 10 iterations.

Global haplotype structure

Here we investigate the impact of global haplotype frequencies on the prediction performance. Figures 2.9 and 2.10 show achieved accuracies for data sets HapMap and BC for all 10 iterations and a range of δ values. The accuracy generally increases with the addition of haplotype frequencies into the prediction. However, the improvement in accuracy varies for different HLA genes and for different iterations. This variability is greater in the HapMap data set, which is likely due to a relatively small sample size considering the high polymorphism of HLA genes. Lower accuracies are likely to be achieved for alleles that are under-represented or non-existent in the training data. In contrast, the BC data set shows less variability in accuracy across different iterations

[8] www-gene.cimr.cam.ac.uk/public_data/HLA/HLA.shtml

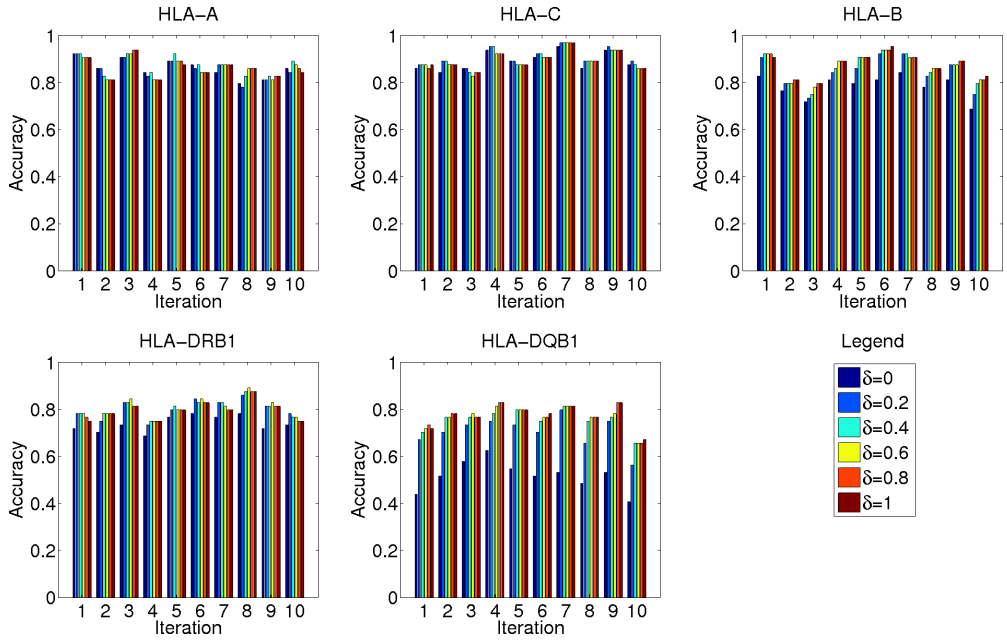

Figure 2.9: Accuracy of prediction for each HLA gene in HapMap data set when using global haplotype frequencies across 10 iterations.

due to its larger sample size. Some variability across different genes is still present; for instance, the HLA-B gene is predicted with the lowest accuracies. This gene is the most polymorphic gene in the region, and therefore, the most challenging one to predict.

In order to evaluate the trend with which the accuracy changes with the addition of haplotype frequencies, we looked at the average increase in prediction accuracy over all iterations, relative to the accuracy for $\delta = 0$ (haplotype information not included in prediction). Accuracies for $\delta > 0$ are divided by accuracy for $\delta = 0$ and averaged over 10 iterations, so that the relative increase in accuracy for $\delta = 0$ for all genes is equal to 1. Relative increase different from 1 indicates the factor by which the accuracy changed from the accuracy at $\delta = 0$. These results are shown in Tables 2.4 and 2.5. The relative accuracy increases for all genes for $\delta > 0$, that is, with the addition of the haplotype structure. This increase is the biggest for HLA-DQB1 (max 1.5102) and smallest for HLA-A (max 1.015) in HapMap data set. Gene HLA-A is farthest away from other HLA genes, and therefore, benefits the least from the addition of haplotype structure. Note that the relative increase in accuracy is generally smaller for the BC data set. This

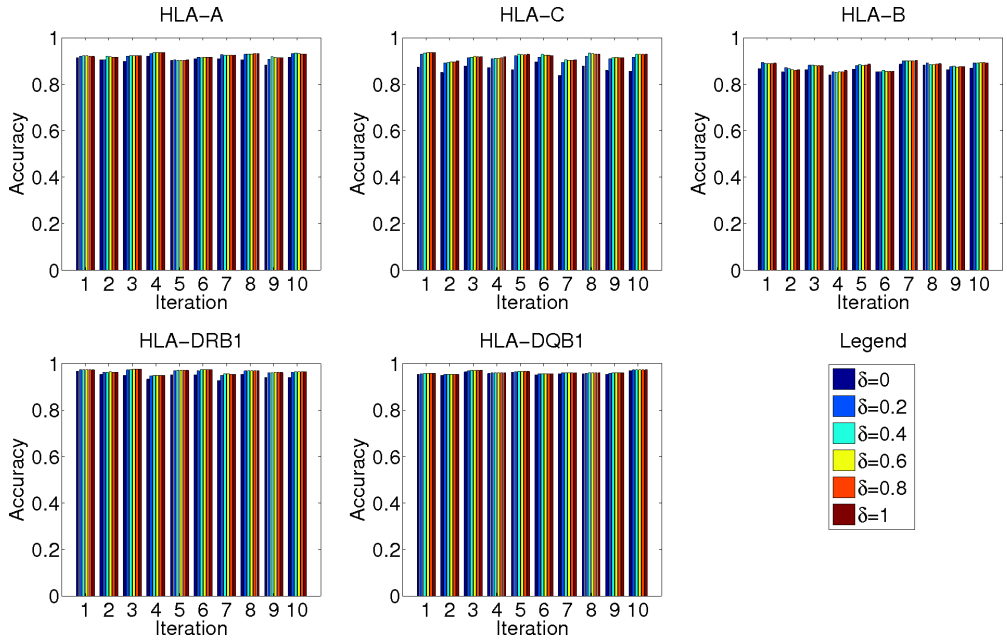

Figure 2.10: Accuracy of prediction for each HLA gene in BC data set when using global haplotype frequencies across 10 iterations.

is due to the overall higher accuracies in this data set (> 0.9), which leaves less room for improvement. It is important to note that there is, however, a steady increase in accuracies as we add the haplotype structure in both data sets, and for all genes.

Table 2.4: Average relative increase in accuracy when using global frequencies in HapMap data set as δ increases

δ	A	C	B	DRB1	DQB1
0	1.0000	1.0000	1.0000	1.0000	1.0000
0.2	0.9963	1.0195	1.0729	1.0846	1.3727
0.4	1.0148	1.0125	1.0955	1.0912	1.4712
0.6	1.0007	1.0020	1.1061	1.0955	1.4917
0.8	1.0025	1.0020	1.1122	1.0787	1.5069
1	0.9989	1.0038	1.1146	1.0765	1.5102

Table 2.5: Average relative increase in accuracy when using global frequencies in BC data set as δ increases

δ	A	C	B	DRB1	DQB1
0	1.0000	1.0000	1.0000	1.0000	1.0000
0.2	1.0145	1.0528	1.0173	1.0174	1.0035
0.4	1.0180	1.0616	1.0166	1.0199	1.0046
0.6	1.0175	1.0616	1.0151	1.0203	1.0046
0.8	1.0168	1.0611	1.0151	1.0199	1.0046
1	1.0171	1.0623	1.0173	1.0199	1.0046

Local haplotype structure

Here we investigate the impact of local haplotype frequencies on the prediction accuracy. Figures 2.11 and 2.12 show the achieved accuracies for data sets HapMap and BC as δ changes from 0 to 1. Note that, in this approach, the prediction for gene HLA-A is independent of the parameter δ and therefore, the accuracy at this gene does not change with δ. Relative increase in accuracy for all genes is shown in Tables 2.6 and 2.7.

Local haplotype frequencies contribute to a steady increase in accuracy across all values of δ with the maximum increase achieved for $\delta = 1$. This is consistent across all genes that are in strong linkage disequilibrium, in this case, genes HLA-B and -C, and HLA-DRB1 and -DQB1. However, when compared against the results obtained by using global haplotype frequencies, we see that comparable accuracies are achieved, for a wider range of δ, rather than for only $\delta = 1$. This is important in practical context of selecting appropriate δ value to predict HLA alleles for any given data set. An additional benefit of using global haplotype frequencies is that the prediction of HLA-A gene is also likely to improve using this approach, unlike the case of using local structure.

Table 2.6: Average relative increase in accuracy when using local frequencies in HapMap data set as δ increases

δ	A	C	B	DRB1	DQB1
0	1.0000	1.0000	1.0000	1.0000	1.0000
0.2	1.0000	1.0000	1.0188	1.0191	1.1431
0.4	1.0000	1.0000	1.0400	1.0473	1.2436
0.6	1.0000	1.0054	1.0497	1.0709	1.3908
0.8	1.0000	1.0197	1.0748	1.0663	1.5237
1	1.0000	1.0184	1.1038	1.0750	1.5496

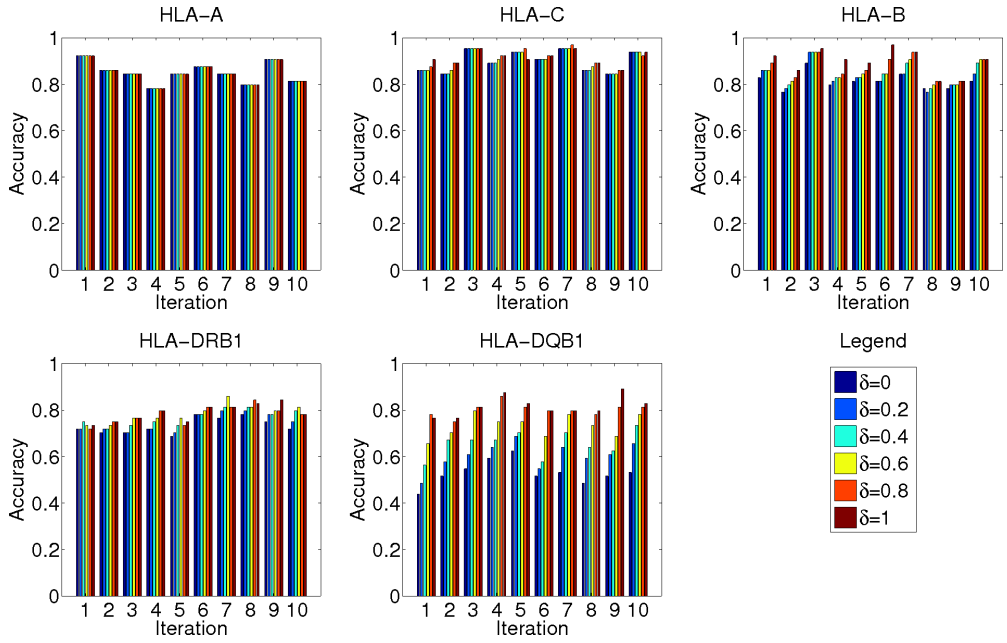

Figure 2.11: Accuracy of prediction for each HLA gene in HapMap data set when using local haplotype frequencies across 10 iterations.

Table 2.7: Average relative increase in accuracy when using local frequencies in BC data set as δ increases

δ	A	C	B	DRB1	DQB1
0	1.0000	1.0000	1.0000	1.0000	1.0000
0.2	1.0000	1.0088	1.0046	1.0011	1.0004
0.4	1.0000	1.0235	1.0069	1.0047	1.0007
0.6	1.0000	1.0484	1.0118	1.0094	1.0018
0.8	1.0000	1.0727	1.0181	1.0150	1.0024
1	1.0000	1.0676	1.0046	1.0196	1.0044

2.7 Conclusion

We demonstrate the need to use high linkage disequilibrium in the MHC region to assist the prediction of HLA alleles. We do this by using haplotype frequencies to show that adding the frequency information increases the accuracy of prediction. In addition, we described an improvement of our original method that allows for a more systematic

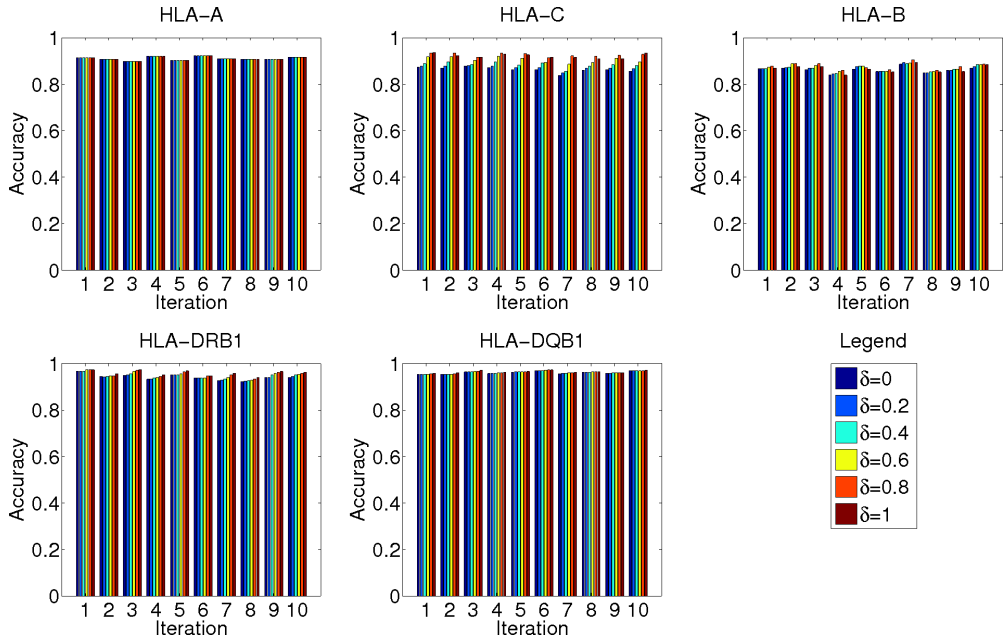

Figure 2.12: Accuracy of prediction for each HLA gene in BC data set when using local haplotype frequencies across 10 iterations.

selection of top predictions. Next, we evaluated the impact of HLA gene dependency on the prediction of HLA genes from SNP data. We proposed integrating local and global dependencies among HLA genes into the prediction, and analyzed the advantages and disadvantages of each.

The results on our data show that the addition of global structure produces a more robust prediction with respect to parameter δ. Furthermore, adding global haplotype information improves the prediction of all genes, including the ones that are farther away from strongly linked local haplotypes, such as gene HLA-A. However, considering the relatively small sample size of our data, we are hesitant to generalize our findings on other data sets. The local structure approach may have a bigger impact on larger data sets that have a better representation of rare alleles. Local linkage between genes contributes with higher local frequencies to the final prediction score and may help promote those rare alleles to the top from the initial list of predictions. Also, linkage patterns differ across populations, and it is widely observed that linkage disequilibrium in non-African populations extends over longer distances than in Africans [79]. This

means that shorter haplotypes should have a bigger impact in African populations than in non-African populations. Since we conducted our study on a data set of European ancestry, we cannot generalize to data sets of other origins.

We plan to extend our work in the following directions:

1. We intend to conduct a thorough experimental evaluation of the proposed methodology. This includes investigation on how experimental parameters impact the overall performance of the method. Some of these parameters are: the number of top predictions k, the number of SNPs used in prediction m, and the width of genomic region around each HLA gene from which the SNPs are used in prediction.

2. Our current implementation of the base classifier is computationally expensive and cannot handle large numbers of SNPs. In the future, we plan to use a more scalable classifier that will handle larger numbers of SNPs and therefore make better initial predictions of HLA alleles.

3. We use a weighted sum function to combine allele and haplotype scores into a final score. We intend to investigate several other functions and their impact on prediction performance of the algorithm.

4. We intend to integrate this framework with a more sophisticated method for gene-independent HLA prediction that produces somewhat higher accuracies.

5. We plan to investigate the application of this framework on a data set of non-European origin. As demonstrated on the BC data set, there is little room for improvement on large data sets of European origin, since high accuracies are obtained even without the addition of HLA haplotype information. However, data sets of non-European origin are less widely available and are of much smaller sample size. As such, they present a bigger challenge when it comes to prediction of HLA genes and could benefit from the gene dependency information. In addition, linkage disequilibrium patterns differ in different populations, and the addition of dependency information into the HLA prediction may be even more beneficial for a non-European population.

6. In this study we used haplotype frequencies published in [37] and generated from a relatively small sample of about $10K$ individuals. In the future we intend to

use more recently published haplotype frequencies generated from a much larger sample size ($\sim 6M$ individuals) and encompass more populations (21 detailed populations) [20]. Using these frequencies may result in higher accuracies as they more adequately represent a population (e.g. using African Black population frequencies published in [20] may enhance predictions for YRI population used in this study because they match the YRI population more closely than African American frequencies published in [37]).

Chapter 3

Measuring Ambiguity in HLA Typing Methods

3.1 Introduction

The high polymorphism in HLA presents a challenge when it comes to typing HLA genes. The typing has historically been performed using serological antibody tests, which are able to identify HLA protein variants on the surface of the cell using antigen-specific antibodies [9]. Serology has been widely replaced with DNA-based typing methods due to its inability to identify all specific products of the HLA alleles. It is shown that in order to improve the clinical outcome of HSCT from an unrelated donor, it is essential to identify and match patient's and donor's HLA genes at the allele level [46, 57, 66].

DNA-based methods identify HLA alleles by interrogating the nuclear DNA sequence and can result in different levels of ambiguity depending on typing methodology or test kits used. HLA typing methods, their corresponding formats and abbreviations used in this chapter are given in Table 3.1. Some of the widely used molecular methods for typing HLA genes, as defined in American Society for Histocompatibility and Immunogenetics Standards (ASHI), are a nucleic acid-based typing method using sequence-specific oligonucleotide hybridization (SSO) [25, 48], a nucleic acid amplification-based typing method using sequence-specific priming (SSP) and sequence-based typing (SBT) [65]. Even though DNA-based technology improved identification of specific alleles, HLA typing results reported by testing laboratories are still commonly resolved to a

certain level of ambiguity, rather than to an exact allele assignment [23]. Exact high resolution HLA typing can be costly and laborious with the large and rapidly growing number of described HLA alleles, which sometimes cannot be easily distinguished with existing high-throughput typing methods. Ambiguous allele assignments are produced either due to failure to interrogate all polymorphic positions, or due to a lack of phase between polymorphisms within a locus because of diploid sequence reads (or both).

Typing Method	**Description**			**Abbr.**	**Example Typing**
Serology	Identifies HLA protein on the cell surface using antigen-specific antisera	Broad antigen typing		SERO	A9, A28
		Split antigen typing			A24, A68
DNA	Identifies HLA alleles by interrogating DNA	Allele family level	Two-digit resolution	DNA2	A*24:XX, A*68:XX
		Sequence specific oligonucleotides	Allele codes	SSO	A*24:AER, A*68:GM AER=02/03/04/05 GM=01/02/03/04/05
		Sequence based typing	Single-pass SBT	SBT	A*24:02, A*68:01 or A*24:03, A*68:01 or A*24:04, A*68:01 or A*24:05, A*68:01
		High resolution	Exact alleles	HR	A*24:02, A*68:01

Table 3.1: HLA typing formats reported by NMDP contract typing laboratories.

In HSCT the selection of donors for a patient in need of a transplant is based primarily on HLA matching, and the lower the ambiguity of typing the easier it is to determine the probability of allele level match during the donor search [26]. In order

to facilitate rapid identification of matched donors for HSCT several methods have been proposed to infer unknown phase and allele assignment [19, 35]. They typically employ statistical methods and the unique properties of the HLA region, such as its high linkage disequilibrium, in order to estimate haplotype frequencies and predict the most likely haplotype assignment for an individual with an HLA typing consisting of a set of ambiguous allele pairs. HLA typing with less ambiguity on average gives fewer high-probability phased high resolution haplotypes. We aim to measure per-locus ambiguity resulting from several HLA typing formats across four continental populations using an information theory-based measure, Shannon's entropy [71].

Some previous work has been done in measuring typing ambiguity, first a measure developed by Helmberg et al. [22] and more recently, the first application of Shannon's entropy to this problem by Cano [8]. Helmberg proposed a characterization of HLA typing kits using a frequency inferred typing (FIT) index. A FIT index describes the probability of correct allele pair assignment for an ambiguous typing result, and is calculated as the negative log of the probability of a wrong allele pair prediction. This probability is equal to the sum of products of all allele pairs that share the same typing pattern as the selected pair. The limitation of the FIT index is that it does not take into account the distribution of allele frequencies beyond the most likely assignment. The concept of measuring ambiguity in HLA typing using entropy was first presented by Cano in [8]. They used population-specific allele frequencies and several SSO typing examples to demonstrate the utility of the measure.

Both previous typing ambiguity studies used allele frequencies in their computations and, as we will later show, fail to demonstrate the advantage of linkage disequilibrium information contained in haplotype frequencies when it comes to reducing ambiguity and improving predictions of patient-donor matching. We use haplotype frequencies and show that the ambiguity is reduced considerably compared to using allele frequencies, proving that this advance in strategy for identifying matched donors has had a significant positive impact. In addition, we take this methodology further and use it to evaluate several different typing methods, and directly compare them with respect to the inherent ambiguity measured by entropy. To measure the impact of the typing method ambiguity in more relatable terms, we also developed a measure we call Confirmatory Typing (CT) Mismatch Rate, which gives the average probability across a

set of patients that a mismatch would occur between the patient and donor when a high resolution confirmatory typing is performed on the ambiguously-typed donors in a uniformly-typed registry.

We show that entropy can be used to objectively compare methods of HLA typing to each other in terms of the information they provide, in the context of each individual population. Our results show that intermediate-resolution single-pass sequence-based typing (SBT) reported in genotype list format contains the least ambiguity and, therefore, the most certainty in allele prediction across all populations. We examine the benefit of using haplotype frequencies in entropy calculations versus allele frequencies. Neighboring HLA and non-HLA genes are highly correlated and major efforts have been directed at describing linkage disequilibrium (LD) across the region [11, 44, 80]. When certain alleles occur together generally due to linkage disequilibrium between them, some ambiguity can be inferred away using this linkage information, which is inherently contained in haplotype frequencies. Our results show that using population haplotype frequencies immensely reduces the ambiguity present in HLA typing. This demonstration allows HLA typing methods to be objectively evaluated in the practical context of a matching algorithm that uses haplotype frequencies to predict probabilities of allele level matches between a patient and list of potentially matched donors. It is hoped that this analysis can lead to data-driven HLA typing resolution strategies for registry donor and cord-blood unit (CBU) typing.

3.2 Materials and Methods

3.2.1 Typing formats

Before DNA-based HLA typing methods were developed, serological testing identified sets of alleles with similar reactivity. Two-digit level resolution is the lowest HLA typing resolution reported by typing laboratories today. For these lower-resolution formats, alleles in the same family as A*01:01 are reported as A1 using serological methods (abbreviated SERO), or as a truncated result of intermediate-level DNA-based typing (referred to as DNA2 in this text) as A*01 or A*01:XX.

A commonly used intermediate-resolution format is the one using NMDP allele codes [5]. Sequence-specific oligonucleotides (SSO) typing results are reported in this format

for this study, where each allele code represents two or more alleles. For example, an allele reported as A*01:AB can be either A*01:01 or A*01:02, and an allele reported as A*26:JGSJ can be any of the following three: A*25:13, A*26:01, A*26:52. Therefore, the number of combination for an ambiguous allele pair reported in this format increases multiplicatively, that is, the allele pair (A*01:AB, A*26:JGSJ) will have six possible pairwise combinations (A*01:01, A*25:13 or A*01:01, A*26:01 or A*01:01, A*26:52 or A*01:02, A*25:13 or A*01:02, A*26:01 or A*01:02, A*26:52). Ambiguous sequence based typing (SBT) is reported in the format of genotype lists for this study, that is, in the form of several possibilities for pairs of alleles an individual carries (A*24:02,A*68:01 or A*24:03,A*68:01 or A*24:04,A*68:01). Because in single-pass SBT results, some ambiguous genotype lists cross several allele families, allele codes could be used to represent all typing results. However, genotype list representation has the advantage of showing that some genotype possibilities, added implicitly when compressing to allele code format, are not possible.

3.3 Data Sets

3.3.1 Haplotype frequency data

We used high-resolution haplotype frequencies generated from unrelated donors from the National Marrow Donor Program (NMDP) database for four principal population categories defined by the United States census: African American (AFA), Caucasian (CAU), Hispanic (HIS) and Asian/Pacific Islander (API) [37]. These categories are referred to as self-described ethnic groups (SIRE), as they are selected by individuals from the NMDP race/ethnicity questionnaire at donor registration. NMDP develops and maintains a repository of several million HLA-typed donors to facilitate hematopoietic stem cell transplantations among unrelated individuals. Table 3.2 shows populations used and the number of haplotypes, HLA-A, -B, and DRB1 alleles within each population.

3.3.2 Simulated typing results

To generate simulated typings for different HLA typing methods, we first sampled 2 haplotypes from a high resolution population haplotype frequency data set [37]. These

Population	Description	# of 3-locus Haplotypes	# of HLA-A Alleles	# of HLA-B Alleles	# of HLA-DRB1 Alleles
AFA	African American	3,049	68	107	59
CAU	Caucasian	5,214	97	158	70
API	Asian-Pacific Islander	2,157	56	102	62
HIS	Hispanic	3,102	75	138	62

Table 3.2: **HLA haplotype frequency data used in this study.** This table shows four population groups and their corresponding haplotype frequencies used for the simulation of samples in this study. The data contains frequencies for three-locus haplotypes (A B DRB1). The table also shows the number of unique HLA-A, HLA-B and HLA-DRB1 alleles present in the haplotypes for each population group.

sampled haplotypes were then "rolled up" from the high resolution typing to a lower resolution typing to emulate how the typing would have appeared using various typing methods. For example, a high resolution haplotype pair with the format:

$$A*23:01 \sim B*18:01 \sim DRB1*07:01, A*30:02 \sim B*58:02 \sim DRB1*12:01$$

would be rolled up into lower resolution typing (in this case serology) as follows:

$$A23, A30; B18, B58; DR7, DR12.$$

Note that the high-resolution haplotypes contain neither phase nor allele ambiguity, while the lower resolution typing contains both. Simulated typings of 1,000 individuals were generated for the four broad population groups (AFA, API, CAU, HIS) and four different typing methods (SERO, DNA2, SSO, SBT).

While we use HLA nomenclature Version 3 style formatting to describe HLA alleles in this chapter, we simulated the four typing methods for Version 2.28 of the IMGT-HLA database, to more closely match the time in which the typing results used to generate the haplotype frequencies were reported. To generate serologic typing (SERO), we used the HLA dictionary, which allows each HLA allele to be mapped to a serologic equivalent (e.g. B*15:02 maps to B75) [24]. To map alleles to DNA 2-digit (DNA2), we

removed all fields from the HLA typing but the first field describing the allele family (e.g. B*15:02 becomes B*15:XX). To simulate SSO typing, given detailed information on the probes present in each SSO kit and the IMGT/HLA database of allele sequences, probe hit tables were computed for all possible combinations of described alleles. Each pair of alleles was mapped to the set of allele pairs that had identical probe hit patterns, then the typing was compressed to NMDP allele code format. For HLA-A and HLA-B loci we used kits described at the 12th International Histocompatibility Workshop [31], and for HLA-DRB1 locus we used a kit described at the 11th International Histocompatibility Workshop [48, 47]. SBT simulation mapped allele pairs to a list of ambiguous genotypes with identical heterozygous sequence in exons 2 and 3 published by IMGT-HLA (http://www.ebi.ac.uk/imgt/hla/ambig.html) and reported the typing in the genotype list format.

3.3.3 Shannon's entropy

Shannon's entropy quantifies the amount of uncertainty or disorder associated with a particular system, and is widely used in a variety of applications, such as genetics [30, 33, 83], data mining [51], molecular biology [67], and imaging [69]. In information theory, entropy is used to measure uncertainty associated with a random variable. Here we used entropy to measure and compare ambiguity associated with the results of various HLA typing methods. Given an ambiguous genotype X, Shannon's entropy (H) is defined as:

$$H(X) = -\sum_{x \in X} p(x) log(p(x)) \tag{3.3.1}$$

Where $p(x)$ is the relative frequency of a single-locus or multi-locus high resolution genotype x. Entropy can be thought of as the ambiguity or impurity present in a system of interest. If, in a set of typing results for one individual, all of them are equally likely (frequent) then the entropy is the highest as we have the least information to choose the most likely real genotype.

To illustrate the usefulness of entropy in measuring typing ambiguity we show an example of two typing results with the same number of ambiguities (Table 3.3). Both of these ambiguous typings have six possible pairs of alleles, however, one of them is more

pure with respect to the frequency distribution of these combinations, and therefore has a lower entropy (0.014 versus 1.676, with a mean entropy of 0.54 for ambiguous typing results in this sample). Note that entropy depends on the size of the set as well as the distribution of frequencies in the set. If H_n is the entropy of n typings, then it is maximal for outcomes of equal frequencies, that is, $H_n(p_1, .., p_n) \leqslant H_n(1/n, .., 1/n)$, and it increases with the number of equally frequent typings, that is, $H_n(1/n, .., 1/n) < H_{n+1}(1/(n+1), .., 1/(n+1))$. Therefore, depending on the distribution of frequencies and the number of possible alleles for an ambiguous allele pair, typing results can have drastically different entropies.

Typing 1			Typing 2		
Ambiguities	Relative Frequency	$-plog(p)$	Ambiguities	Relative Frequency	$-plog(p)$
B*5702/B*5801	0.0923	0.3172	DRB1*0301/DRB1*1301	0.9989	0.0016
B*5702/B*5802	0.0832	0.2985	DRB1*0301/DRB1*1327	0.0005	0.0056
B*5702/B*5804	0.0001	0.0016	DRB1*0304/DRB1*1301	0.0002	0.0024
B*5703/B*5801	0.4331	0.5229	DRB1*0304/DRB1*1327	0.00	0.00
B*5703/B*5802	0.3907	0.5297	DRB1*0306/DRB1*1301	0.0004	0.0044
B*5703/B*5804	0.0006	0.0063	DRB1*0306/DRB1*1327	0.00	0.00
H = 1.676			H = 0.014		

Table 3.3: **An illustration of two ambiguous typing results with the same number of possible allele sub-types and different level of ambiguity as measured by entropy.** Shown here are typing results for two simulated subjects. They each have six ambiguous sub-types (allele-pairs), but very different entropies. Typing 1 has entropy $H = 1.676$ and Typing 2 has entropy $H = 0.014$. Both of these typings come from the same population sample (African American) in which the mean allele entropy for this simulated sample is $H = 0.54$.

3.3.4 Entropy calculations using HLA frequencies

The HLA haplotype frequencies we used in this study were estimated by the expectation-maximization algorithm (EM) described in [29]. To obtain the allele frequency of allele A_j from haplotype frequencies we simply summed the frequencies over all haplotypes in the given population group that included that allele, that is:

$$p(A^j) = \sum_{i=1}^{n} p(A_i \sim B_i \sim D_i) I(A^j) \tag{3.3.2}$$

where A, B, and D are typed loci, N is the total number of haplotypes and $I(A^j)$ is an indicator function that takes value of 1 (0) when a haplotype contains (does not contain) allele A^j. For each ambiguously typed locus, we generated all possible pairs of alleles (A^iA^j), and computed their respective frequencies as follows:

$$p(A^iA^j) = p(A^i)p(A^j). \tag{3.3.3}$$

In the first set of experiments, we used these allele pair frequencies derived from allele frequencies in entropy calculations to compute allele entropy for each locus.

To compute locus entropy using haplotype frequencies, or haplotype entropy, we do the following. For each ambiguously typed three-locus genotype, $g = (AaBbDd)$, we generated all possible haplotype pairs and their frequencies using imputation as described in [29]. To compute the entropy of a particular locus, we obtained frequencies for each unique allele pair on that locus, say A^iA^j, by summing over all frequencies of haplotype pairs generated for the given genotype, that contain the given allele pair, that is

$$p(A^iA^j) = \sum_{k=1}^{N} p(A_k \sim B_k \sim D_k, A_k \sim B_k \sim D_k) I(A^iA^j) \tag{3.3.4}$$

where N is the number of generated haplotypes and $I(A^iA^j)$ is an indicator function that takes value of 1 (0) when a haplotype pair contains (does not contain) allele pair (A^iA^j). We then used these allele pair frequencies derived from haplotype frequencies to compute haplotype entropy in the same manner as described for the case of allele entropy.

As a side note, an ambiguous typing with many possible alleles at each locus can result in a large combinatorial number of possible haplotype pairs. Given a fully heterozygous case of three-locus un-phased genotype with $n_i, i \in 1, 2, 3$ possible alleles at each locus, the number of possible haplotypes is equal to $\prod_{i=1}^{L} n_i$ and the number of phased haplotype pairs is equal to $2^{L-1} \prod_{i=1}^{L} n_i$, where L is the number of loci, in this case $L = 3$. In these equations we assume the heterozygosity of all loci, since the estimated numbers would be smaller for a homozygous case in which some HLA loci have the same alleles on both chromosomes. For typings including five or six HLA loci and high allelic ambiguity, the number of phased haplotype pairs can grow into billions.

3.3.5 Confirmatory typing mismatch rate

Besides objective evaluation of typing methodologies employed in typing the HLA region, using the entropy approach to measure ambiguity has another application from a clinical perspective, namely its direct relationship to confirmatory typing (CT) mismatch rates. For a given patient, high resolution CT is done to confirm the patient-donor match from a selected set of donors. A case where the high resolution typings mismatch is called a CT mismatch. We compute CT mismatch rates on the same simulated donor sample by comparing the ambiguous typing and the exact haplotype pair that was used to generate that ambiguous typing. As described in a previous section, each ambiguous genotype can generate multiple HLA haplotype pairs, the true one being the haplotype pair we used to simulate the ambiguous typing. The CT mismatch rate for each locus is computed as the summation of frequencies of all allele pairs (computed using Equation 3.3.4) that do not match the corresponding allele pair of the haplotypes used to simulate the donor typing. This is the probability that a selected donor will not be the exact match for the given patient.

3.4 Results

Locus entropies obtained for SBT, SSO, allele family level DNA2 and SERO typing methods are shown in Table 3.4, averaged over the three loci (HLA-A, -B, -DRB1) and within each population using allele frequencies (allele entropy). SBT had the lowest entropy and therefore the least inherent ambiguity when it comes to resolving HLA

alleles at the locus level, across all populations. As expected, serology produced the lowest resolution typing and had the highest entropy. SSO typing was far more ambiguous than SBT across all evaluated datasets, reflecting both SBT's more complete coverage of polymorphic positions in the exons and the benefits of using genotype list format for SBT typings rather than the NMDP allele code format we used for SSO. Figure 3.1 shows the average allele entropy for each locus for AFA, CAU, HIS and API population groups. The ranking of the typing methods with respect to the least amount of ambiguity across all populations was: SBT, SSO, allele family level DNA2, SERO.

	AFA	API	CAU	HIS
SBT	0.0354	0.0668	0.0766	0.0787
SSO	0.2974	0.5247	0.49	0.5004
DNA2	1.8501	2.023	1.056	2.1709
SERO	1.7735	2.2282	1.6548	2.9471

Table 3.4: **Average allele entropy for all typing methods and all population groups.** This table shows the locus entropy for SBT, SSO, DNA2 and SERO typing methods for all four populations using allele frequencies and averaged over the three loci, HLA-A, -B, -DRB1.

When we used haplotype instead of allele frequencies, we got the same ambiguity ranking (Table 3.5) for average locus entropies obtained for SBT, SSO, and DNA2 typing methods (haplotype entropy). However, a dramatic decrease in entropy occurred across all typing methods when we used haplotype instead of allele frequencies. For example, the allele entropy of SSO typing in Caucasian group is 0.49 while the haplotype entropy is an order of magnitude lower at 0.0477. This decrease is due to some ambiguity being resolved by LD information provided in haplotype frequencies, and is successfully captured by Shannon's entropy. Figure 3.2 shows the average haplotype entropy for each locus and each population separately. The LD information contained in haplotype frequencies reduces the entropy considerably compared to only using allele frequencies, demonstrating that this strategy for identifying matched unrelated donors has a significant positive impact. Figure 3.3 shows the comparison between allele and haplotype entropies for each typing method and within each population group. This result also demonstrates the utility of imputation algorithms that generate population haplotype frequencies to more accurately predict the likelihood of allele match for stem

cell registry matching algorithms.

	AFA	API	CAU	HIS
SBT	5.30E-04	0.0031	0.002	0.005
SSO	0.0923	0.2074	0.0477	0.1195
DNA2	1.2685	1.6219	0.7277	1.7633
SERO	1.2488	1.3889	0.7205	1.7495

Table 3.5: **Average haplotype entropy for all typing methods and all population groups.** This table shows the locus entropy for SBT, SSO, DNA2 and SERO typing methods for all four populations using haplotype frequencies and averaged over the three loci, HLA-A, -B, -DRB1.

To demonstrate the impact of typing method ambiguity in a clinical setting we computed CT mismatch rates, which give the average probability that a mismatch would occur between a patient and donor when high resolution confirmatory typing is performed on the ambiguously typed donors in a uniformly typed registry. CT mismatch rates computed on the same sample of 1000 simulated donors are shown in Table 3.6. We can see a direct correlation between CT mismatch rates and entropy computed across typing methods dimension (Pearson's correlation coefficient between CT mismatch rates and haplotype entropy is = 0.96, and between CT mismatch rates and allele entropy is = 0.92). For SSO typing we found an average haplotype entropy of 0.05 and a 1.31% CT mismatch rate, while for SBT typing, we found an average haplotype entropy of 0.002 and a 0.08% CT mismatch rate, in the CAU population group. In a hypothetical donor registry with uniformly typed donors, choosing a typing method with smaller entropy across a given population may result in smaller mismatch rates during the confirmatory typing phase and more certainty in a selected set of donors. This CT mismatch rate gives us a more intuitive clinical interpretation of these entropy scores. An important assumption when computing CT mismatch rates is that all donors in the registry are typed with the same method. This is generally not the case, and as donors with better typing accrue, CT mismatch rates are expected to decrease over time.

3.5 Discussion

We have shown that entropy can be used to objectively compare methods of HLA typing in terms of the information they provide. The calculation of per-locus entropy using

	AFA	API	CAU	HIS
SBT	2.0135e-04	9.3340e-04	8.0012e-04	0.0015
SSO	0.0236	0.0530	0.0131	0.0324
DNA2	0.3792	0.4286	0.4069	0.4366
SERO	0.3755	0.3724	0.1957	0.4334

Table 3.6: **Confirmatory typing (CT) mismatch rates for all typing methods and all population groups.** This table shows the expected confirmatory typing (CT) mismatch rates for SBT, SSO, DNA2 and SERO typing formats and four populations averaged across all three loci (HLA-A, -B, DRB1). CT mismatch rates describe the probability that a mismatch would occur between a patient and donor during high resolution confirmatory typing on the ambiguously typed donors in a uniformly typed registry.

haplotype frequencies has a direct application in measuring the impact of using haplotype frequencies to predict the likelihood of allele match for stem cell registry matching algorithms. The LD information contained in haplotype frequencies reduces the entropy considerably compared to using allele frequencies, showing that this strategy for identifying matched donors has a significant positive impact. No objective quantitative comparisons between SBT and SSO methods have been available to date. Typing laboratories may choose SSO methods over SBT methods primarily based on cost savings achieved due to easier set-up, staff training, pre-packaged kits, and automation. However, these apparent cost savings may have a price of higher typing ambiguity. We have shown that single-pass SBT typing performs far better at distinguishing alleles compared to mid-1990's-era SSO typing. However, currently available SSO typing kits used for recruitment typing have more oligonucleotide probes and thus are able to distinguish more alleles, which may result in entropy as low as that of single-pass SBT. Given equal cost, registries should utilize laboratories that employ HLA typing methods that achieve lower entropy for their population. Our objective measure of typing ambiguity can be advantageously applied to the continual improvement of all methods of HLA typing. Design of SSO kits could be done *in silico* using population haplotype frequencies and sequence information from the IMGT-HLA database [62]. SSO kits are designed to distinguish between the most common alleles in a population. In United States and Europe, populations of European origin predominate, and therefore some alleles common in minority populations may not be distinguished in some kits. Given

a fixed number of probes, the SSO kit that provides the lowest entropy in a population could be considered optimal.

As new alleles are discovered, SSO kits are often altered to add more probes so that typing results do not cross allele families and thus meet current guidelines for acceptable recruitment typing. These new probes will not decrease entropy appreciably as the frequency of a newly discovered allele tends to remain very low.

Because of sample size limitations, many of the rare alleles described in IMGT-HLA were not observed in our samples. However, rare alleles do not have a significant impact on entropy calculations. Owing to the logarithmic nature of Shannon's entropy, an allele with a very small frequency p contributes $plog(p)$ to the resulting entropy. This quantity approaches zero for very small values of p. More formally, $\lim_{p\to 0} plog(p) = 0$. This property of the entropy guarantees that potential underestimation of frequencies of rare alleles not included in the population groups in Table 3.2 will only slightly underestimate the typing ambiguity. Larger frequency-generating sample sizes available in the future will serve to eliminate this issue.

Our methods of HLA typing method evaluation can also be applied to next-generation sequencing technologies. Recently the Roche 454 sequencing platform has been employed for HLA typing in research rather than recruitment [4, 73]. The 454 platform has relatively longer read lengths that can clonally type entire exons without intra-exonic phase ambiguity. However, intronic regions are not amplified by this platform, thus the system lacks the inability to phase across exons leading to some remaining genotypic ambiguity, which would be reflected in entropy calculations. Meanwhile, other next-generation sequencing platforms more commonly used for whole genome sequencing use a shotgun approach for sequence coverage that includes intronic regions of HLA genes [68]. The Illumina platform uses short reads and high read depth, and there is a potential for HLA typing ambiguity to vary between sequencing runs on the same sample because of differences in read coverage, and thus success with assembly [81]. With the ambiguity of current SBT methods caused by the reading of heterozygous sequence from two chromosomes simultaneously, a future technology that would allow for a single chromosome to be read clonally could eliminate haplotype ambiguity [56]. It is important to note that many genome-wide studies in practice do not make HLA allele calls because the polymorphism of the HLA system requires specialized bioinformatics

analysis unique to these genes [49].

A consideration specifically related to the HLA typing method is the representation of the ambiguous allele data derived from the HLA typing, which was also measured using entropy. The 2-digit DNA typing resolution is in practice a result of incomplete reporting of SSO, SSP, or SBT typing data. The higher entropy of this type of data shows the value in reporting the complete information available from the HLA typing platform rather than rounding to the allele family level. The genotype list representation yields a slightly lower entropy than the NMDP allele code representation. Genotype list representation allows for the exclusion of some genotypes that have been ruled out by the HLA typing method, but would still be included in the Cartesian product of the alleles listed in the NMDP allele codes.

Note that, in some populations, the HLA-B locus presents higher values of entropy when typed using 2-digit DNA methods than when typed at the serological level (Figure 3.1). This is likely due to the fact that some 2-digit DNA allele families contain alleles from multiple serologically defined categories. For example, serological antigens exist to split alleles in the HLA-B15:XX family into B62, B63, B75, B76 and B77, and HLA-B40:XX family into B60 and B61, while 2-digit DNA typing does not distinguish between them.

This analysis provides a path for defining acceptable HLA typing for recruitment as minimum requirements for entropy scores as a measure of typing ambiguity and for HLA data representation guidelines as a way to ensure that genotype lists are reported. Single-pass or highly automated SBT can result in HLA typings that cross allele families, which does not meet current minimum standards for recruitment typing at NMDP, yet we show that it provides a high-quality low-entropy HLA typing. In fact, we had to use the genotype list representation for the simulation of single-pass SBT typings because some allele combinations result in HLA typings for which no NMDP allele codes have been created due to required minimum standards that HLA allele codes do not generally cross allele families [21]. Requiring laboratories to resolve ambiguous alleles in SBT to meet current NMDP requirements can significantly increase cost, but does not significantly lower entropy. Developing a standard for laboratory reports of typings that cross allele families would thus enable a reduction in cost of recruitment typing without a reduction in quality. We observe variation in entropy for the same HLA typing method

between populations and loci. For example, Figure 3.3 shows lower average entropy of SSO typing results in the CAU sample than in the HIS sample. Some variation is attributed to differences in frequency-generating sample size for a particular population group (in this example simulated Caucasian typings are generated from a larger pool of haplotype frequencies than Hispanic typings, due to a larger availability of Caucasian donors in the registry). The remaining entropy differences can be attributed to the nature and magnitude of HLA genetic diversity in the same group (one can expect HIS population group to be more broadly defined and hence more genetically diverse). The resulting entropy can also depend on LD patterns, the location and number of DNA polymorphisms, and the shape of the allele frequency distribution in the populations.

In addition to absolute differences in entropy between populations, we also observed differences between population groups in the effectiveness of using haplotype frequencies in decreasing haplotype entropy. Having higher levels of LD can improve the predictive capability of haplotype frequencies, and so African population samples with lower LD could have higher entropy than European populations, with higher LD, for this reason. In the opposite direction, higher HLA diversity would lead to higher entropy in African population samples than in European samples. The API sample may have relatively higher entropy than other population samples because the API frequency distribution constitutes an average of the frequency distributions of multiple distinct populations, and thus may be skewed more towards rare types than other populations in this study. If API entropy were evaluated using more detailed race subcategories (e.g. Japanese, Korean, Filipino, etc.), we would expect lower entropy values because the HLA diversity of each respective sub-region would be lower. The size of the population sample used to generate haplotype frequencies also plays a role in the entropy calculations in that a relatively larger sample, as we had for CAU compared to the other races, would give higher entropy. Because of these multiple confounding factors affecting entropy, we urge caution in using entropy as a measure to compare the HLA characteristics between samples of different ancestry. There are some caveats in that the simulation framework implicitly has no sampling error or estimation error in the haplotype frequencies. In practice, uncertainty in the frequency estimates will lead to higher entropy, so our results should be treated as a practical lower bound.

For interpretation of between-locus entropy differences, we turn to the history of

HLA nomenclature in that the allele families and serologic types were defined primarily using European samples. The naming of allele families was based loosely on serologic categories, and at some point in history newly discovered serologic patterns were no longer used to split up allele families. The discovery of new alleles also has an impact on entropy in that some populations have not been well-characterized for HLA and some individuals may have as yet un-described alleles that can result in some hidden entropy. In evaluating entropy at the locus level, we see that at the allele family level, the HLA-DRB1 locus has a higher entropy than HLA-A and HLA-B loci. The number of allele families defined for HLA-A and HLA-B is higher than that of HLA-DRB1, giving a lower entropy for typing resolution at the 2-digit or serologic levels, all else being equal.

Stem cell registries have been accruing HLA typing results for over 25 years, with continual advancement in typing methods during this period. The proportion of donors typed by each method changes over time in a searchable registry due to new donor recruitment, roll-off of donors exceeding the maximum age, reporting of primary HLA data, prospective typing, and high resolution typing on behalf of patients. With analysis of changes in HLA typing data for each donor over their time on the registry, it becomes possible to chart decreasing entropy in HLA typing over time and determine which typing methods were primarily responsible for this decrease. Entropy could also be applied as a selection factor for prospective typing projects in which some donors are upgraded to lower ambiguity typings.

In summary, the application of Shannon's entropy as a measure of HLA typing ambiguity has benefits throughout the lifecycle of HLA typing: in reagent design, lab reporting standards, donor recruitment typing guidelines, and registry matching algorithm performance evaluation.

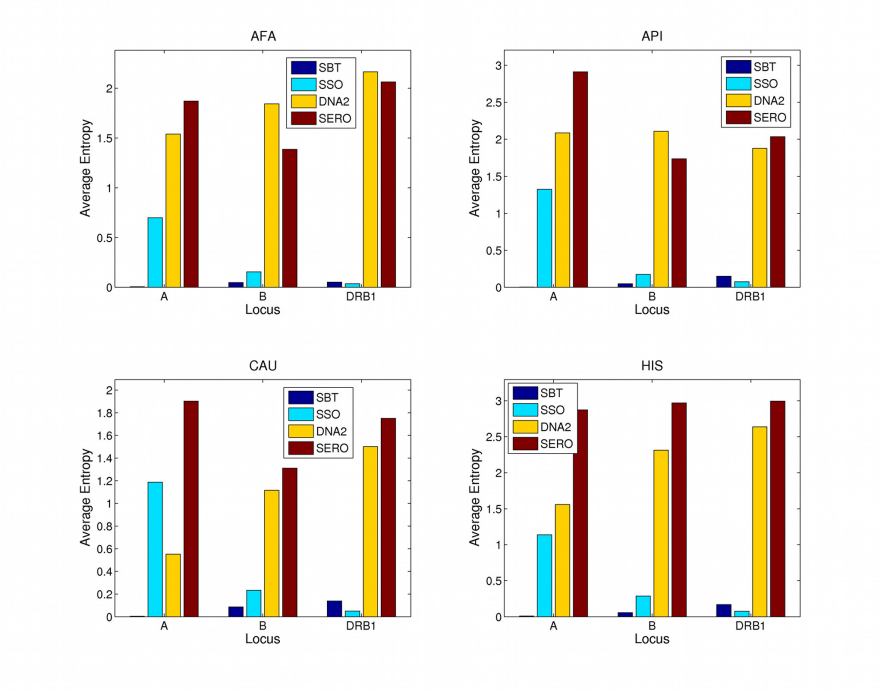

Figure 3.1: **Average allele entropy.** This figure shows locus entropies obtained for SBT, SSO, DNA2 and SERO typing formats, within each population and for three HLA loci using allele frequencies, that is, the allele entropy. The four panels correspond to four populations: AFA (African American), API (Asian-Pacific Islander), CAU (Caucasian) and HIS (Hispanic), respectively, from left to right, top to bottom. The y-axis shows entropy averaged across 1000 simulated donor typings, and the x-axis corresponds to the HLA locus that the entropy is measured for (HLA-A, HLA-B and HLA-DRB1). The color represents the typing methods used for typing: SBT (single pass sequence-based typing), SSO (sequence-specific oligonucleotides), DNA2 (two-digit allele family level DNA-based typing) and SERO (serological typing).

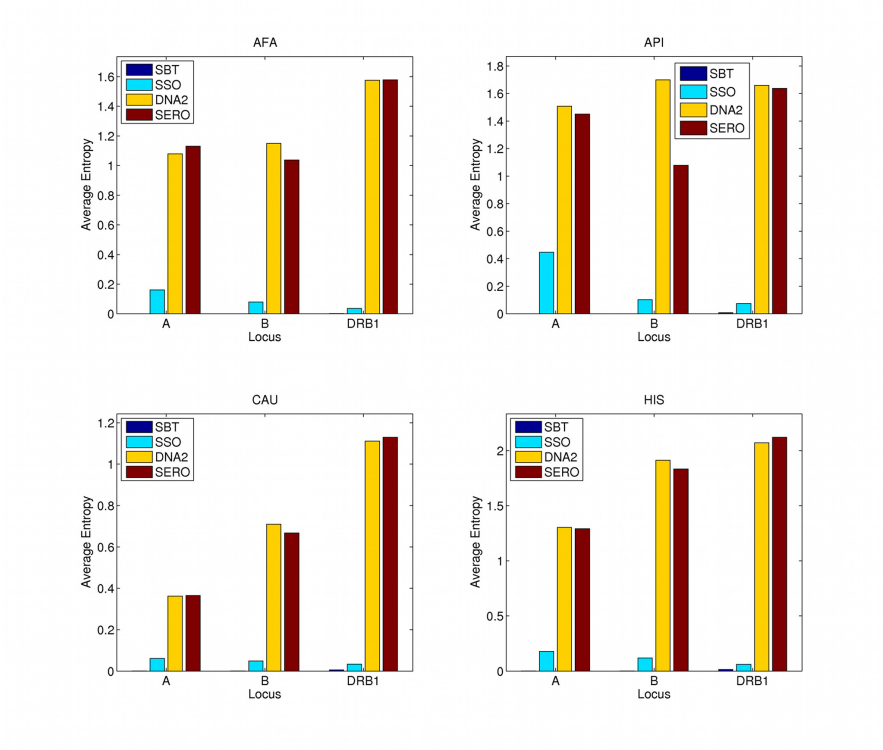

Figure 3.2: **Average haplotype entropy.** This figure shows locus entropies obtained for SBT, SSO, DNA2 and SERO typing formats, within each population and for three HLA loci using haplotype frequencies, that is, the haplotype entropy. The four panels correspond to four populations: AFA (African American), API (Asian-Pacific Islander), CAU (Caucasian) and HIS (Hispanic), respectively, from left to right, top to bottom. The y-axis shows entropy averaged across 1000 simulated donor typings, and the x-axis corresponds to the HLA locus that the entropy is measured for (HLA-A, HLA-B and HLA-DRB1). The color represents the typing methods used for typing: SBT (single pass sequence-based typing), SSO (sequence-specific oligonucleotides), DNA2 (two-digit allele family level DNA-based typing) and SERO (serological typing).

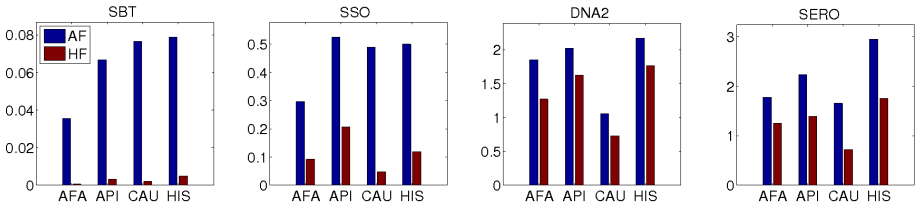

Figure 3.3: **Comparison of average per-locus entropies obtained from allele and haplotype frequencies.** This figure shows the comparison between allele entropy and haplotype entropy computed for four typing methods: SBT (single pass sequence-based typing), SSO (sequence-specific oligonucleotides), DNA2 (two-digit allele family level DNA-based typing) and SERO (serological typing), respectively from left to right. Allele entropy was computed using allele frequencies (AF) and genotype entropy was computed using haplotype frequencies (HF). The y-axis shows the average entropy values, and the x-axis shows the four continental populations for which the entropy was computed: AFA (African American), API (Asian-Pacific Islander), CAU (Caucasian) and HIS (Hispanic), respectively. In all figures, the locus entropy is averaged across the three loci, HLA-A, -B, -DRB1.

Chapter 4

SNP-based Prediction From Ambiguous HLA Data

4.1 Introduction

The typing of HLA alleles is generally performed using DNA-based assays that are not always able to precisely identify the alleles present in the tested sample. Exact allele-level HLA typing can be costly and laborious with the large and rapidly growing number of described HLA alleles. Some of the widely used molecular methods for typing HLA genes are sequence-specific oligonucleotide (SSO) hybridization, sequence-specific primer (SSP) amplification and sequence-based typing (SBT) [52].

While the typing methodology evolved over time and will evolve further, the majority of a stem-cell registry data still consists of ambiguous HLA assignments [20]. Ambiguity at a single HLA gene comes in two forms: *allele ambiguity*, where the polymorphisms that distinguish alleles are not interrogated by the typing system, and *phase ambiguity*, which results from the inability to establish chromosomal phase between identified polymorphisms. All HLA data obtained using current typing methodologies comes without phase, that is, contains phase ambiguity. The Be The Match registry contains HLA typing of approximately 10 million donors as of this year. All volunteer donors were typed for at least HLA-A and -B when they joined the Registry. Most donors were typed using serologic methods until 1997, when an initiative to use DNA-based testing was implemented [29]. Starting in 1992 many new volunteers were also typed at the

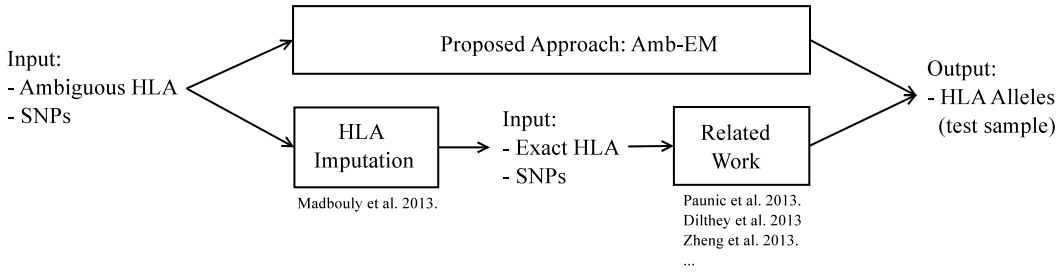

Figure 4.1: SNP-based prediction of HLA alleles: proposed approach and related work.

HLA-DRB1 gene. Over time the HLA DNA-based testing has incorporated additional genes, and the resolution has increased from low to intermediate and high. Even though the typing quality has increased for new donors, in order to utilize the huge amounts of historical HLA data in the Registry operations or research, we must develop algorithms that can successfully learn from ambiguous data.

The described *allele* and *phase* ambiguity found in most HLA data sets today presents a challenge to current SNP-based methods for prediction of HLA alleles, since they require non-ambiguous data. A way to obtain high-resolution HLA alleles from ambiguous data is to use one of the existing methods for resolving ambiguity from the ambiguous HLA data [36] and use the generated high-resolution HLA alleles in learning the SNP-based prediction method (Figure 4.1). However, depending on the levels of ambiguity in the HLA data, the errors generated by these methods can be prohibitively high, up to 45% (this error includes the error in both, allele and phase assignment, and should be interpreted as the upper bound on the allele assignment error) [36]. These errors are then further propagated when the incorrectly imputed HLA alleles are used to train the SNP-based prediction model.

In our previous work we proposed a method to infer HLA alleles from SNP genotypes and population HLA haplotype frequencies [53, 54]. We used an expectation-maximization (EM) based approach to assign the most likely HLA assignment to a new test sample. In this chapter we propose an approach that utilizes ambiguous HLA data in SNP-based prediction. The contributions of this chapter are the following:

(a) We propose a novel approach to SNP-based prediction of HLA alleles that handles not only *phase* ambiguity, but also *allele* ambiguity in the HLA genotypes, and thus

uses ambiguous HLA data as a learning source in the prediction (Figure 4.1).

(b) Additionally, we evaluate how different levels of ambiguity in HLA data impact the prediction performance.

(c) Finally, we demonstrate that building the predictive model from ambiguous, rather than from statistically imputed high-resolution data, is a more accurate approach to SNP-based prediction of HLA alleles.

4.2 Methods: *Amb-EM* Algorithm

Here we describe *Amb-EM*, a novel approach to predicting HLA alleles from SNP data by using EM to resolve *allele* and *phase* ambiguity in the training data. This approach builds upon our previous work described in [54]. The novelties of *Amb-EM* are described in Section 4.2.1, while Section 4.2.2 is briefly summarized from [54].

4.2.1 Step 1: Independent EM classifiers

Classifiers are trained for each gene separately using SNPs within and around that gene, that is, to classify HLA-A alleles, we use SNPs within and around HLA-A only, and so on. We first use the EM to model the joint distribution of SNP and HLA alleles by estimating the frequency of their joint (SNP, HLA) haplotypes, and then use Bayes' rule to compute the probability of all pairs of HLA alleles given a new sample with SNP genotype data.

Training

Given ambiguous HLA and SNP genotypes for all samples in the training data set, EM estimates the frequencies of all haplotypes that could be obtained from the data. The outcome is a set of joint (SNP, HLA) haplotypes h and their estimated frequencies f. To do this, the EM algorithm expands the un-phased genotype data into all possible phased pairs of haplotypes. For instance, given four SNPs $s = (AT, GG, CT, TT)$, all possible pairs of haplotypes that could be obtained from this data are:

$$s = (AT, GG, CT, TT) \quad \rightarrow \quad \begin{array}{l} h_1 = A\ G\ C\ T, h_2 = T\ G\ T\ T \\ h_1 = A\ G\ T\ T, h_2 = T\ G\ C\ T \end{array}$$

where h_1 and h_2 are the two haplotypes.

In this example, the SNP genotype data only contains phase ambiguity, that is, the exact allele assignments are known at each SNP. In addition to phase ambiguity, HLA data often contains allele ambiguity as well. For example, an individual's typing at gene HLA-A may result in two ambiguous alleles g_1 and g_2, where g_1 is either $A*01:01$ or $A*01:02$ and g_2 is one of the following three alleles: $A*25:13$, $A*26:01$, $A*26:52$. All possible HLA-A genotypes are therefore all possible combinations of g_1 and g_2 ambiguities, that is

$$g = (g_1; g_2) = (A*01:01, A*01:02; \quad A*25:13, A*26:01, A*26:52) \rightarrow \begin{array}{r} A*01:01, A*25:13 \\ A*01:01, A*26:01 \\ A*01:01, A*26:52 \\ A*01:02, A*25:13 \\ A*01:02, A*26:01 \\ A*01:02, A*26:52 \end{array}$$

where g is ambiguous HLA typing at gene $HLA-A$. A comma is used to separate HLA allelic ambiguities, while a semi-colon is used to separate a pair of HLA alleles on the two chromosomes.

With each HLA ambiguity, the number of possible phased pairs of (SNP, HLA) haplotypes increases exponentially. For example, given three SNPs, $s = (AT, GG, CT)$, and two ambiguities at gene $HLA-A$, $g = (A*01:01, A*01:02; \quad A*26:01, A*26:52)$, the number of possible (SNP, HLA) haplotype pairs is $n = 16$, that is

$$(s,g) = (AT, GG, CT, A{*}01:01, A{*}01:02; A{*}26:01, A{*}26:52) \rightarrow \begin{array}{l} A\ G\ C\ A*01:01, T\ G\ T\ A*26:01 \\ A\ G\ C\ A*01:01, T\ G\ T\ A*26:52 \\ A\ G\ C\ A*01:02, T\ G\ T\ A*26:01 \\ A\ G\ C\ A*01:02, T\ G\ T\ A*26:52 \\ T\ G\ T\ A*01:01, A\ G\ C\ A*26:01 \\ T\ G\ T\ A*01:01, A\ G\ C\ A*26:52 \\ T\ G\ T\ A*01:02, A\ G\ C\ A*26:01 \\ T\ G\ T\ A*01:02, A\ G\ C\ A*26:52 \\ A\ G\ T\ A*01:01, T\ G\ C\ A*26:01 \\ A\ G\ T\ A*01:01, T\ G\ C\ A*26:52 \\ A\ G\ T\ A*01:02, T\ G\ C\ A*26:01 \\ A\ G\ T\ A*01:02, T\ G\ C\ A*26:52 \\ T\ G\ C\ A*01:01, A\ G\ T\ A*26:01 \\ T\ G\ C\ A*01:01, A\ G\ T\ A*26:52 \\ T\ G\ C\ A*01:02, A\ G\ T\ A*26:01 \\ T\ G\ C\ A*01:02, A\ G\ T\ A*26:52 \end{array}$$

This number is a power of 2 larger than the number of haplotype pairs in the case of no ambiguities at the gene A. For example, for the same three SNPs and non-ambiguous HLA-A alleles, $g = (A*01:01; \quad A*26:01)$, the number of possible haplotype pairs is $n = 4$, that is

$$(s,g) = (AT; GG; CT; A*01:01, A*26:01) \rightarrow \begin{array}{l} A\ G\ C\ A*01:01, T\ G\ T\ A*26:01 \\ T\ G\ T\ A*01:01, A\ G\ C\ A*26:01 \\ A\ G\ T\ A*01:01, T\ G\ C\ A*26:01 \\ T\ G\ C\ A*01:01, A\ G\ T\ A*26:01 \end{array}$$

Now, given a sequence of un-phased SNP genotypes and ambiguous un-phased HLA genotypes, EM expands the sequence into all possible pairs of haplotypes and collects all unique (SNP, HLA) haplotypes. It then iterates between pairs of (SNP, HLA) haplotypes and a set of (SNP, HLA) haplotypes observable from the data, in the following manner:

- *Expectation* - uses current haplotype frequencies to calculate conditional probabilities of each possible pair of haplotypes given the observed SNP and HLA data for each subject. These are then used to update current frequencies of each pair of haplotypes.

- *Maximization* - uses the frequencies of pairs of haplotypes from the previous step to update haplotype frequencies. In the initial step, haplotype frequencies are set to $1/n_h$ where n_h is the number of unique haplotypes in the data set.

These two steps are repeated until estimated frequencies do not change or the pre-set number of iterations is exceeded. The outcome is a set of joint (SNP, HLA) haplotypes h that agree with the observed genotype data in the training sample and their estimated frequencies $f(g_a, s_{a_1}, s_{a_2}, ..., s_{a_m})$ where s_a and g_a denote a SNP and HLA allele respectively, and m is the number of SNPs.

Prediction

Let $g_i = (g_{i1}, g_{i2})$ denote an HLA genotype of subject i and $s_i = (s_{ij}, \forall j \in [1, m])$ represent a SNP genotype. Let s_a denote a SNP allele, and g_a an HLA allele. Given haplotype frequencies $f(g_a, s_{a_1}, s_{a_2}, ..., s_{a_m})$ estimated in the previous step by EM, and a new sample with a SNP genotype $s' = (s'_j, \forall j \in [1, m])$, the probability that this sample has any two HLA alleles $g = (g_1, g_2)$ can be calculated as follows:

$$P(g|s') \propto P(g, s') \tag{4.2.1}$$

where $P(g, s') = P(g_1, g_2, s')$ is a joint probability of HLA and SNP genotypes, and can be obtained by summing the frequencies over the set H of all (HLA, SNP) haplotypes that can be generated from that genotype:

$$P(g, s') = P(g_1, g_2, s_1, s_2, ..., s_m) = \sum_H f(g_1, s_{a_{11}}, s_{a_{21}}, ..., s_{a_{m1}}) * f(g_2, s_{a_{12}}, s_{a_{22}}, ..., s_{a_{m2}}). \tag{4.2.2}$$

The top predicted HLA allele pair for a new SNP genotype s' is that for which the probability $P(g|s')$ is maximal:

$$(g'_1, g'_2) = \arg \max_{(g_1, g_2)} P((g_1, g_2)|s'). \tag{4.2.3}$$

An example of expanding a new SNP genotype s' and an allele pair $g = (A*01 : 01, A*02 : 50)$ into all possible pairs of haplotypes and computing their joint probability is shown in Chapter 2 in Figure 2.5.

We assign a prediction score to each of the two selected alleles to obtain an allele score as follows:

$$S_{g'_1} = S_{g'_2} = P((g_1, g_2)|s') \tag{4.2.4}$$

We use these scores along with HLA haplotype frequencies in gene-dependent prediction described next.

4.2.2 Step 2: Correlated prediction using HLA haplotype frequencies

Top predictions

Instead of keeping only the first allele pair predicted by EM classifiers for each HLA gene, we keep several top predictions by sorting them in descending order of their prediction likelihood. We then move down the list until the cumulative prediction likelihood exceeds a prespecified threshold, and keep the visited predictions. We obtain top predicted allele pairs for each sample in the testing data and for all HLA genes independently.

HLA haplotype frequencies

Given a set of likely predictions at each gene, we use population haplotype frequencies [20] to narrow down the selection of HLA alleles and make final predictions. The haplotype frequencies are estimated at the level of the entire US population and inform us of how likely sets of alleles are to co-occur in each separate population [20]. We utilize the structure that exists between the genes as an additional source of information in order to improve the prediction.

We do this in the following manner: given a set of selected allele pairs for each gene, we generate all possible pairs of HLA haplotypes. Let's call this set H. Given a haplotype frequency table constructed from the population haplotype frequencies, we retrieve the frequency of each HLA haplotype from H and assign a haplotype score to each pair of haplotypes by multiplying their respective frequencies.

To select the most likely pair of haplotypes for a test sample with unknown HLA alleles, we combine the allele prediction score obtained by EM, S_A, and the haplotype score obtained in this step, S_H, into a weighted score S as follows:

$$S = (1 - \delta) * S_A + \delta * S_H \tag{4.2.5}$$

where $\delta \in [0, 1]$. The predicted pair of haplotypes for a new sample is the one with the highest score S. For small δ, predictions are based more heavily on independently obtained allele score S_A and less on haplotype score S_H, and vice versa.
The entire method is shown in Algorithm 2.

Algorithm 2 *Amb-EM*

for each new sample s **do**
  Construct modified EM classifier for each gene
  Find top predictions for s for each gene
  Assign prediction likelihoods as allele scores of top allele pairs
  Using top predictions, generate all possible pairs of HLA haplotypes
  Retrieve haplotype frequency for candidate haplotypes
  Compute a score that combines allele and haplotype scores
  Determine a pair of haplotypes such that the overall score is maximized
end for

4.3 Results and Discussion

4.3.1 Experimental goals

We conducted several experiments to evaluate the performance of our proposed approach on a real world data set. More specifically, we explored the following research questions:

- Is it better to incorporate *allelic* ambiguity straight into the model or to resolve the ambiguity through statistical imputation before building the model? We investigated this question by comparing the results obtained by running *Amb-EM* on the ambiguous data and the results obtained by running our previous method that does not handle ambiguity on data imputed to high-resolution. (Section 4.3.3).

- How do different amounts of ambiguity in the training data impact the prediction performance? We varied the level of ambiguity in the training data and compared the obtained accuracies for each ambiguity level (Section 4.3.4).

We use accuracy, as the percentage of correctly predicted alleles, as the evaluation metric in all experiments. Section 4.3.2 describes the data sets used in this study.

4.3.2 Data sets

We used HLA and SNP data from the 1000 Genomes project (KG)[9]. This data set contains 930 individuals as summarized in Table 4.1. The 1000 Genomes individuals are typed at $10,268$ SNPs in the MHC region and 5 HLA genes (HLA-A, -C, -B, -DRB1, -DQB1).

[9] http://www.1000genomes.org/

Abbreviation	Population	Count	Broad	Count
CEPH	CEPH individuals	45		
FIN	HapMap Finnish individuals from Finland	93		
GBR	British individuals from England and Scotland	89		
TSI	Toscan individuals	90	Europe (EUR)	317
CHB+JPT	Han Chinese in Beijing, Japanese individuals	165		
CHS	Han Chinese South	100	Asia (ASI)	265
CLM	Colombian in Medellin, Colombia	60		
MXL	HapMap Mexicans from LA California	55		
PUR	Puerto Rican in Puerto Rico	55	Americas (AM)	170
ASW	HapMap African individuals from SW US	53		
LWK	Luhya individuals	87		
YRI	Yoruba individuals	38	Africa (AFR)	178
			Total	**930**

Table 4.1: The 1000 Genome (KG) data set.

We also used SNP and HLA genotype data from the British 1958 Birth Cohort (BC)[10]. Genotyping of the BC data was carried out using Illumina Human1M-Duo [7]. High resolution typing of five HLA genes (HLA-A, -C, -B, -DRB1, -DQB1) is obtained using Sequence Specific Oligonucleotide methodology[11]. After processing, this data set to include samples that have all 5 HLA genes and a high SNP overlap with the 1000 Genomes data set we, we ended up with 100 samples of European origin with all 5 HLA genes. There are 3267 SNPs that overlap between the two data sets.

[10] www.b58cgene.sgul.ac.uk

[11] www-gene.cimr.cam.ac.uk/public_data/HLA/HLA.shtml

4.3.3 Prediction on ambiguous and imputed data

Here we compare the results obtained by running *Amb-EM* on ambiguous data with the results obtained by running our previous method described in Chapter 2 on imputed high-resolution data. We aim to evaluate whether *Amb-EM*, which incorporates uncertainty in HLA data straight into the SNP-based HLA prediction model, outperforms the model built from imputed HLA data. We trained the prediction model on a randomly selected 80% subset of the data and tested it on the remaining 20%. Table 4.2 shows the overall accuracy of prediction using *Amb-EM* for each HLA gene, across all populations, as well as the accuracy by population. Highest accuracy is achieved on the European populations (EUR), while the worst accuracy is achieved for African (AFR) and Hispanic (AM) populations. This result agrees with previous findings and the expectations, since HLA genes in European populations are on average less genetically diverse than in other populations, while in African populations, they are more genetically diverse, and therefore more difficult to predict.

Table 4.2: Accuracy at 2-field resolution for all available samples with ambiguities.

$\delta = 0.5$	Overall	AFR	ASI	EUR	AM
A	0.75	0.61	0.65	0.83	0.79
C	0.88	0.87	0.82	0.90	0.88
B	0.74	0.61	0.76	0.84	0.58
DRB1	0.73	0.56	0.77	0.83	0.56
DQB1	0.91	0.89	0.91	0.90	0.94

Table 4.3: Accuracy at 2-field resolution for all available samples with no ambiguities (data previously imputed to high resolution).

$\delta = 0.5$	Overall	AFR	ASI	EUR	AM
A	0.74	0.60	0.65	0.83	0.76
C	0.86	0.87	0.85	0.89	0.81
B	0.73	0.60	0.76	0.82	0.58
DRB1	0.73	0.57	0.76	0.83	0.55
DQB1	0.91	0.88	0.91	0.90	0.94

Next, we resolved the ambiguity in the training data through imputation [36] and ran our previous method described in Chapter 2 on the resolved data. The results of this experiment are shown in Table 4.3. On average the prediction results obtained

by *Amb-EM* on ambiguous data are better than on statistically imputed data. Figure 4.2 shows the percentage change in accuracy when using ambiguous rather than the imputed HLA data. The height of the bars indicate the improvement in accuracy of the *Amb-EM* on ambiguous data over the accuracy of the old method on the imputed data. Overall, the change in accuracy is positive when we incorporate HLA ambiguity in the prediction model. For example, in AM populations, the accuracy at gene HLA-C is increased by as much as 7%. Incorporating the HLA ambiguity into the prediction process avoids the error that comes from imputing high-resolution HLA alleles from ambiguous HLA data, and using them in the prediction process.

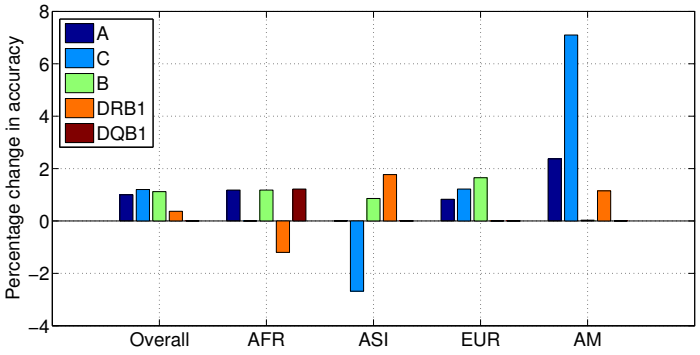

Figure 4.2: Percentage change in accuracy when using *Amb-EM* on ambiguous data over the accuracy of the old method [54] on the imputed data.

The results, however, were not conclusive. We ran a 5-fold cross-validation on the same data, where we divided the data set into 5 folds, using in turn, 1 out of 5 folds for testing and the remaining folds for training. Figure 4.3 shows median (blue bars) and 25^{th} and 75^{th} percentiles (red lines) of the change in accuracy measured in percentages. While the median change in accuracy for genes HLA-A, HLA-C and HLA-B is positive, the change in accuracy for genes HLA-DRB1 and HLA-DQB1 is negative and equal to zero, respectively. One of the reasons for the the poor performance at these two genes is that our data set contains relatively little ambiguity at these two genes. Gene HLA-DRB1 has almost no uncertainty across the data set, with the average number of ambiguous alleles equal to 1.08 per sample. Gene HLA-DQB1 has slightly higher average number of ambiguous alleles (equal to 1.92), but still lower than the number of

ambiguous alleles for other genes. With such high resolution data for these two genes, the HLA imputation algorithm most likely performs well and predicts correct alleles. On the other hand, training *Amb-EM* on ambiguous data introduces some error, which is reflected in poorer performance of our method.

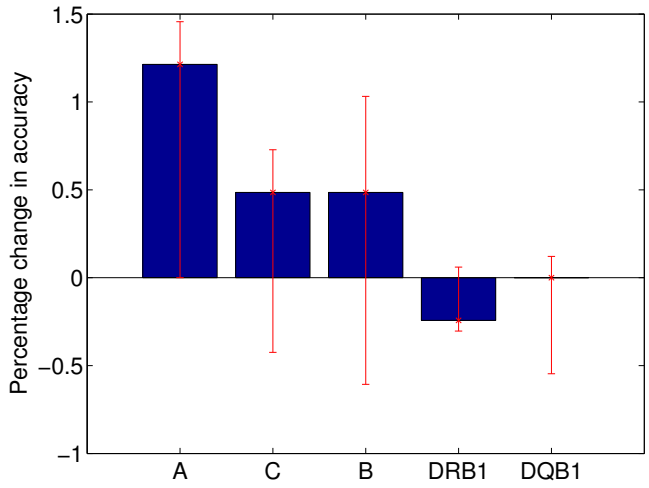

Figure 4.3: Percentage change in accuracy when using *Amb-EM* on ambiguous data over the accuracy of the old method [54] on the imputed data using 5-fold cross validation.

Another thing to note is that each population in our data set is represented by a relatively low sample size. To adequately represent the existing genetic diversity in the HLA region, many more samples are needed. The increased sample size is likely to be beneficial to the SNP-based prediction of HLA alleles, while it does not change the result of the statistical HLA imputation, as it is a deterministic process. This means that, in the case of an increased sample size, the accuracy of prediction from ambiguous HLA data and SNPs will increase, while the imputed HLA data will remain the same, and the increased sample size will have no effect on consequent SNP-based prediction. To demonstrate the sample size effect on our method we randomly selected 80% of the data for training and 20% for testing. We then fixed the test samples, and varied the size of training, by randomly sampling 1/3, 2/3, and 3/3 of the training data for training three different models. Figure 4.4 demonstrates that as we increase the training sample size the accuracy of prediction improves. The difference in performance of using ambiguous

data directly in training rather then imputing it first is, therefore, likely to be even more striking on a larger data set.

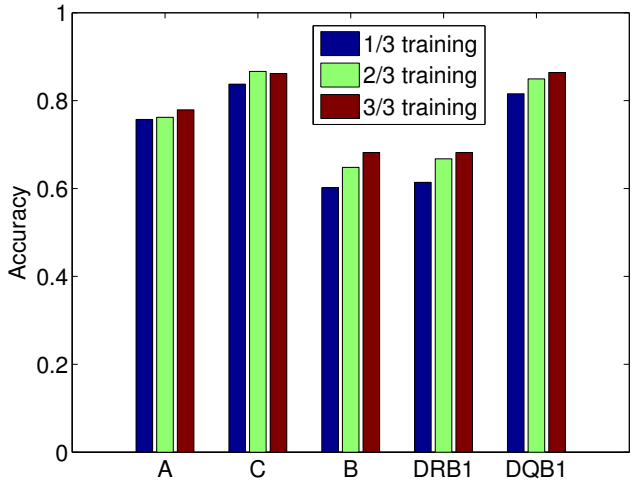

Figure 4.4: The effect of sample size on the prediction performance of *Amb-EM*

4.3.4 Prediction from different levels of ambiguity in training data

In this section we investigate how different levels of ambiguity in the training data impact the prediction performance of *Amb-EM*. We vary the level of ambiguity in the training data and calculate the accuracy on the test data. Due to computational demands of running the experiments on the entire data set, we used a subset of the data of European ancestry for this evaluation (317 European samples from the KG data set and 100 BC samples).

We created three different data sets by selecting 20% of the data for testing and 80% for training in the following manner:

- HR-train (high-resolution training) - samples with low allele ambiguity across all genes are selected for training
- R-train (random training) - data randomly split into training and testing
- LR-train (low-resolution training) - samples with high allele ambiguity across all genes are selected for training

Table 4.4: Accuracy at 2-field resolution for different levels of ambiguity in training data.

	LR-train	**R-train**	**HR-train**
A	0.801	0.867	0.952
C	0.705	0.849	0.934
B	0.705	0.849	0.843
DRB1	0.801	0.819	0.747
DQB1	0.843	0.910	0.916

The obtained accuracies at 2-field resolution data and for $\delta = 0.5$ for all three data sets are shown in Table 4.4. In general the highest accuracy is obtained for the HR-train data set, that is, the data set with the lowest level of ambiguity in the training data. This is expected, since the uncertainty in the training data introduces some error into the prediction model. Figure 4.5 shows that the biggest increase in accuracy as the ambiguity in training decreases is seen at genes HLA-A and -C. However, for other genes, we do not see a clear increasing trend in accuracy as the ambiguity in training decreases. This is due to the varying levels of ambiguity across different genes, because the training and testing data for the three scenarios are selected based on the overall number of ambiguous genotypes across all genes, rather than individual genes.

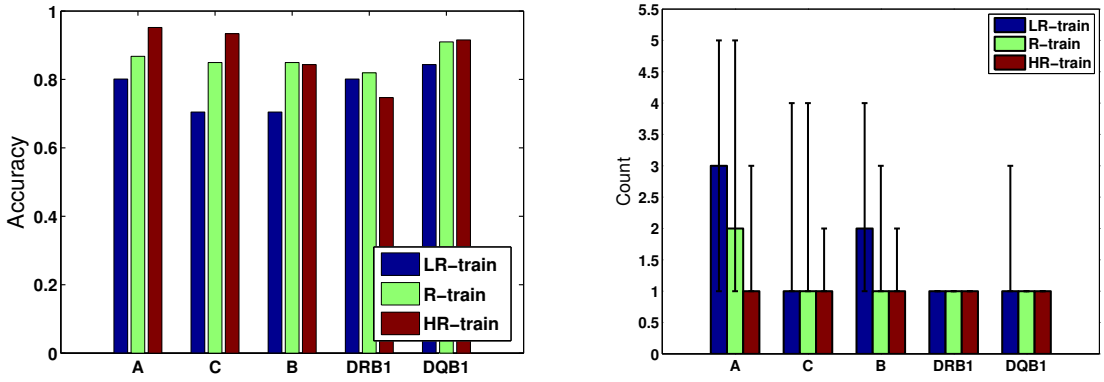

Figure 4.5: Figure on the **left** shows the accuracy achieved on the test data for different levels of ambiguity in the training data. The Figure on the **right** shows the count of ambiguous genotypes in the training data for all three cases: HR-train, R-train, and LR-train.

For example, there is less difference in ambiguity levels within genes HLA-DRB1 and -DQB1 in the three scenarios, than within genes HLA-A and -C. This is demonstrated in the Figure 4.5, where the right plot shows the count of ambiguous genotypes in the training data for each gene. The height of the bars corresponds to the median and the lines mark the 25^{th} and 75^{th} percentile of the count. Notice that genes HLA-A, -C, and -B contain the most ambiguity in the training, while there are very few ambiguous genotypes at genes HLA-DRB1 and -DQB1. Gene HLA-DRB1 also contains ambiguous samples but they do not fall into the 75^{th} percentile. Given similar levels of ambiguity at genes DRB1 and DQB1 in the three scenarios, the variation in accuracy comes from the allele variation in the training and testing samples. This result also demonstrates that for low levels of ambiguity the loss in prediction accuracy is not significant and that the performance of *Amb-EM* is comparable to the performance of our previous method on imputed high-resolution HLA data.

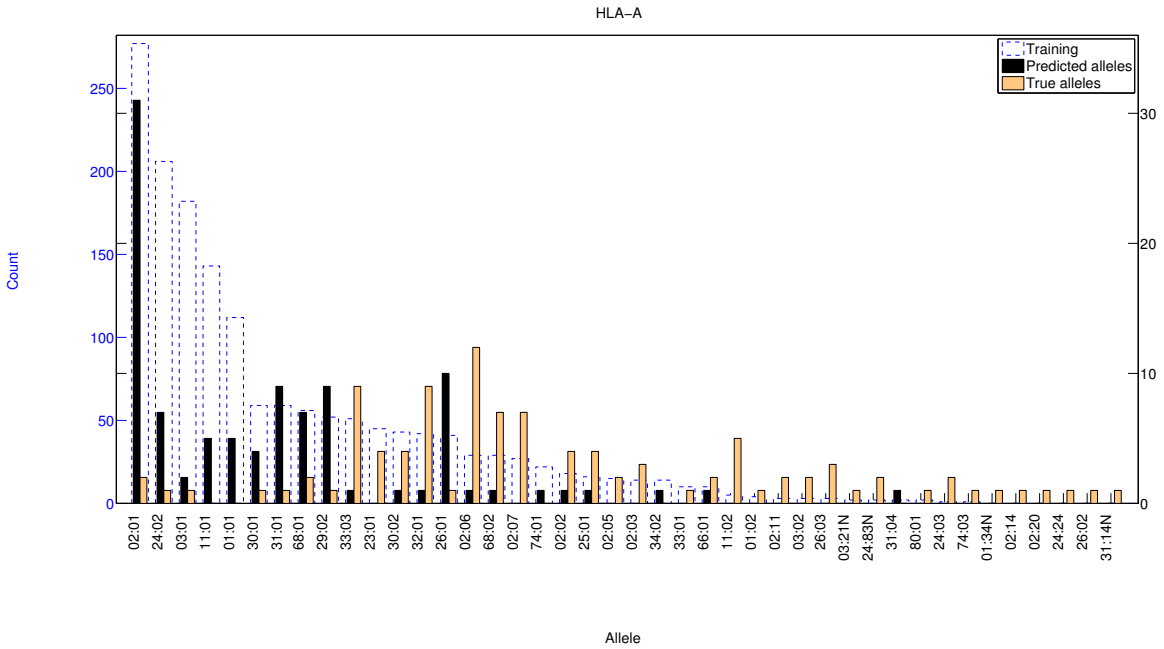

Figure 4.6: Analysis of errors for HLA-A alleles. Blue bars show a count of alleles in the training data set (left y-axis). Black and brown bars represent the count of errors (incorrect predictions) and the count of true alleles that were incorrectly predicted in the test data set, respectively (right y-axis). Note that left and right y-axis have different scales.

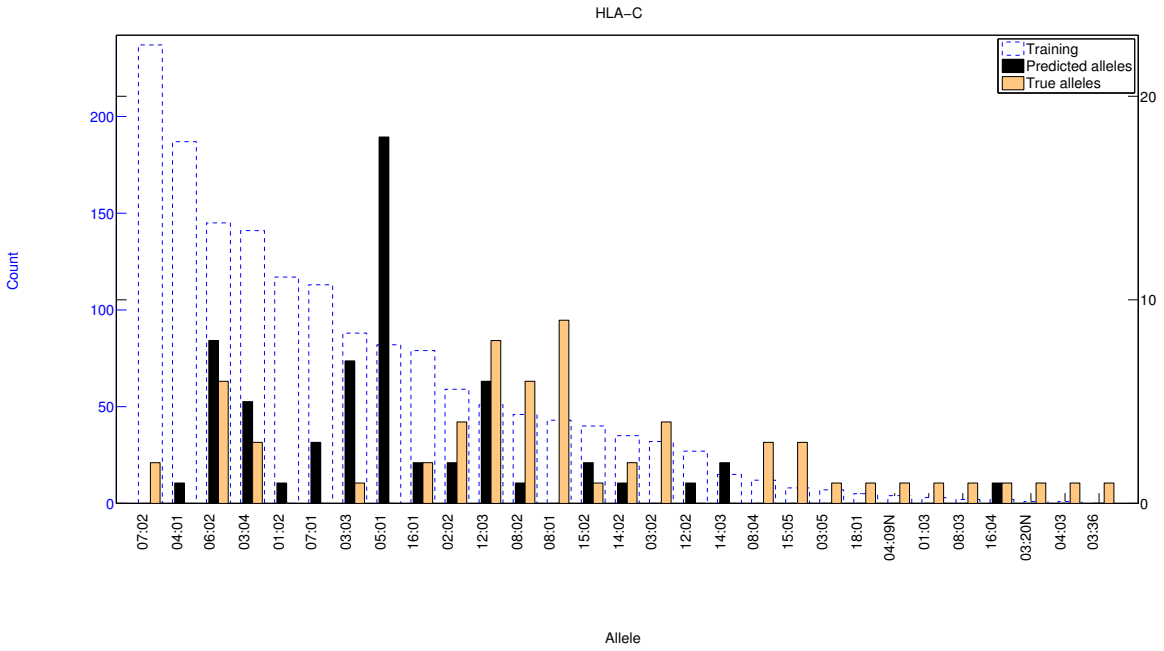

Figure 4.7: Analysis of errors for HLA-C alleles. Blue bars show a count of alleles in the training data set (left y-axis). Black and brown bars represent the count of errors (incorrect predictions) and the count of true alleles that were incorrectly predicted in the test data set, respectively (right y-axis). Note that left and right y-axis have different scales.

Additionally, we took a closer look at the errors predicted by our method in order to identify factors influencing its performance. Since the HLA alleles are not uniformly distributed in a population, some alleles may be more difficult to predict than the others. More specifically, rare alleles that occur with a very low frequency in a sample will be more challenging to predict than the ones that occur very frequently. We showed this in our data by looking at the distribution of the errors predicted by our method and their occurrence in the training sample. We randomly selected 80% of the data for training and 20% of the data for testing. Figures 4.6, 4.7, 4.8, 4.9, and 4.10 show alleles that were incorrectly predicted for HLA-A, -C, -B, -DRB1, and DQB1 genes, respectively, and their occurrence count in the training sample. Blue bars in the figures represent allele frequencies in the training data set. For a given allele: black bars represent those alleles in the test set that were predicted to be the given allele but are not, while brown bars represent those alleles in the test set whose type is that of the given allele that were

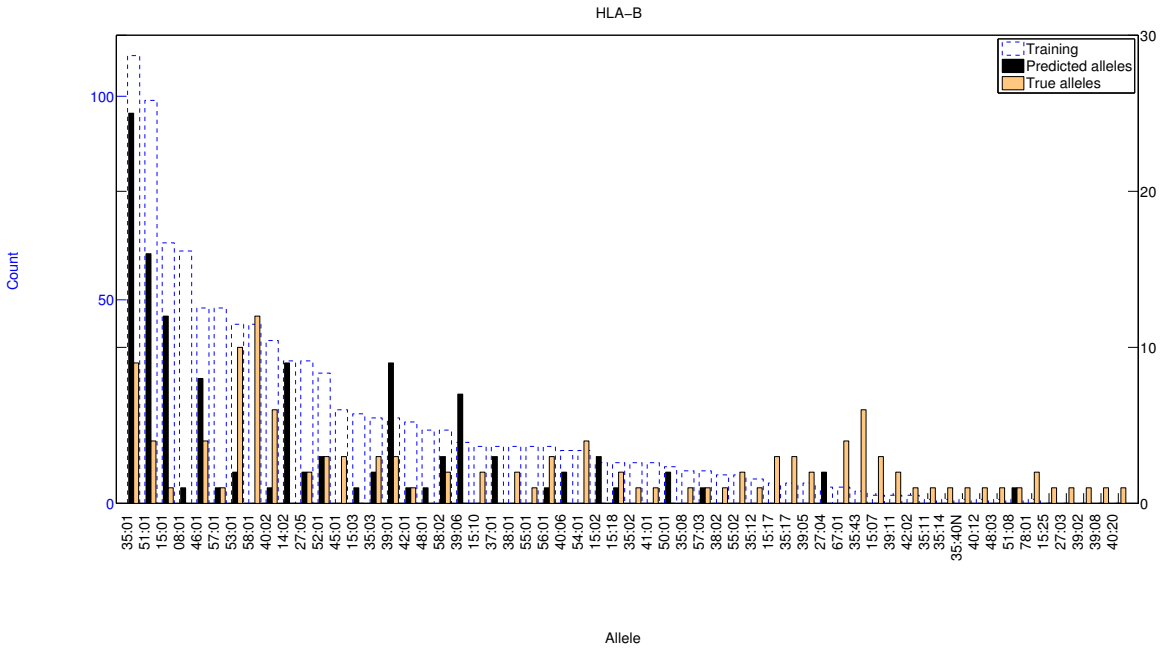

Figure 4.8: Analysis of errors for HLA-B alleles. Blue bars show a count of alleles in the training data set (left y-axis). Black and brown bars represent the count of errors (incorrect predictions) and the count of true alleles that were incorrectly predicted in the test data set, respectively (right y-axis). Note that left and right y-axis have different scales.

predicted to be a different allele. For example, in Figure 4.7, there are two HLA-C*07:02 alleles in the testing set that were incorrectly predicted as different alleles, and there is one allele in the testing set that is predicted as HLA-C*04:01 but is truly a different allele. The alleles are sorted in the descending order of their frequency in the training sample.

In general, the lower the frequency of an allele in the training data set, the more likely it is to be incorrectly predicted if found in the test data set. For example, alleles HLA-A*24:24 (Figure 4.6) and HLA-C*03:36 (Figure 4.7) do not occur at all in the training sample, and therefore, cannot be correctly predicted in the test sample. Other alleles, like HLA-DRB1*14:05 (Figure 4.9) and HLA-DQB1*03:04 (Figure 4.10) occur in the training sample, but with a low frequency. On the other hand, there are alleles that occur with a relatively high frequency in the training data and still occur as errors. These incorrect predictions could be attributed to the population diversity in our data.

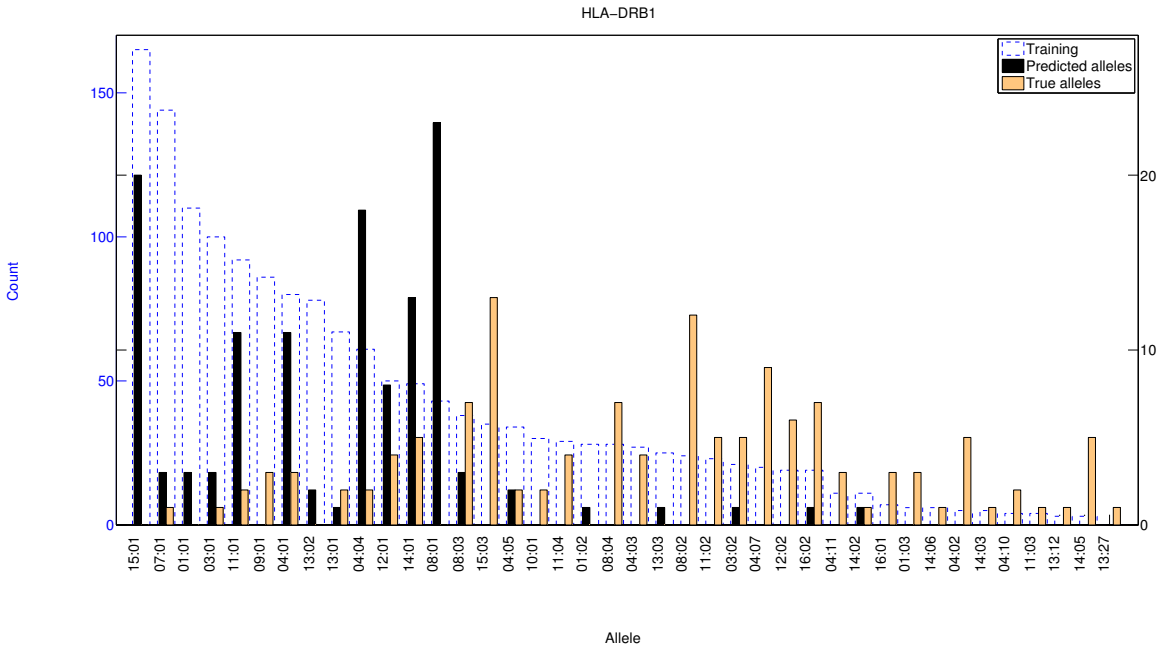

Figure 4.9: Analysis of errors for HLA-DRB1 alleles. Blue bars show a count of alleles in the training data set (left y-axis). Black and brown bars represent the count of errors (incorrect predictions) and the count of true alleles that were incorrectly predicted in the test data set, respectively (right y-axis). Note that left and right y-axis have different scales.

We used all of the data in this evaluation (KG and BC samples) in order to increase the sample size. However, building separate models for each distinct population in the sample may be a better approach and a way to avoid errors of this type. Another impact of the population heterogeneity can be seen around the incorrect predictions that are most often made (black bars in the figures). For example, most of the true alleles that are incorrectly predicted for HLA-A gene are predicted into the allele HLA-A*02:01. This allele has the highest frequency for European populations[12] (Figure 4.6). The same is true for alleles HLA-C*05:01, HLA-B*35:01, HLA-DRB1*08:01, HLA-DQB1*06:02 (Figure 4.7, 4.8, 4.9, 4.10, respectively). As European sample is the largest in our data (417 total samples in KG and BC data sets) the incorrect predictions tend to gravitate towards alleles most frequent in this population.

We also discovered, as a result of this evaluation, that incorrectly predicted alleles

[12] http://www.pypop.org/popdata/

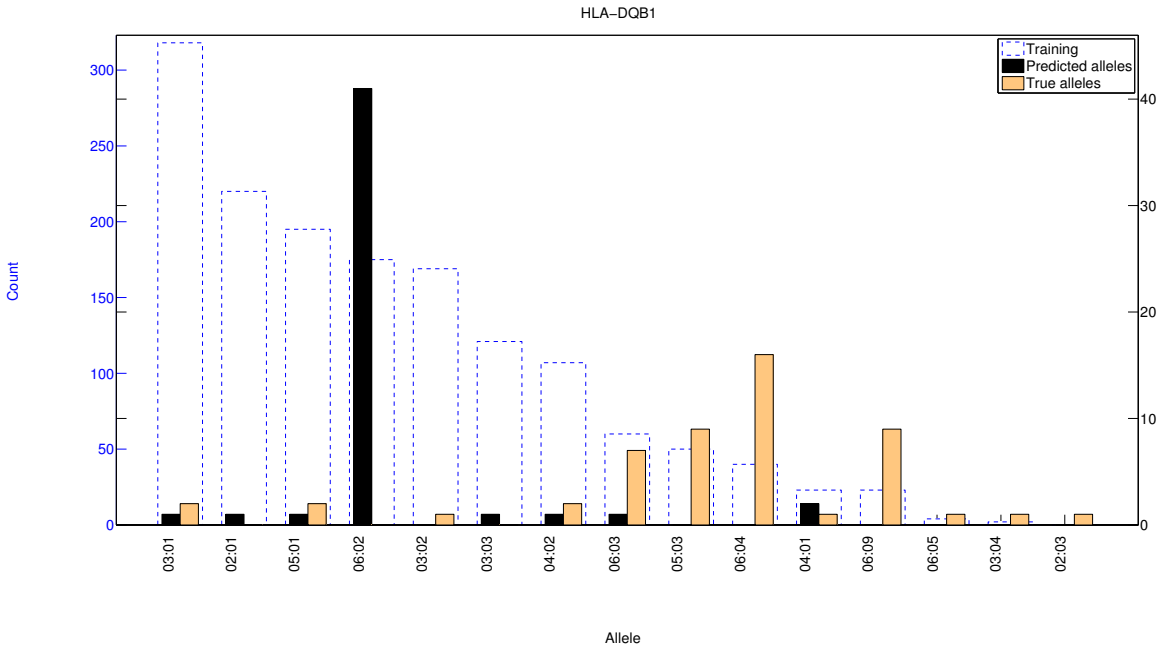

Figure 4.10: Analysis of errors for HLA-DQB1 alleles. Blue bars show a count of alleles in the training data set (left y-axis). Black and brown bars represent the count of errors (incorrect predictions) and the count of true alleles that were incorrectly predicted in the test data set, respectively (right y-axis). Note that left and right y-axis have different scales.

tend to get predicted into the same allele family (1-field group). For example, all samples with incorrectly predicted allele HLA-A*02:01 (black bars in Figure 4.8) have one of the following true alleles: HLA-A*02:02, A*02:03, A*02:05, A*02:06, A*02:07, A*02:11, A*02:14, A*02:20. Similarly, all samples with the incorrectly predicted allele HLA-A*68:01 have A*68:02 as a true allele. Some of the reasons for this type of error is either the lack of the SNPs to distinguish between the different alleles within the same allele family, or the inability of our method to capture those differences.

4.4 Conclusion

In this chapter we proposed a SNP-based approach for prediction of HLA alleles that utilizes ambiguous HLA genotypes as a learning source when building the predictive

model. Additionally, we evaluated how different levels of ambiguity in HLA data impact the prediction performance and concluded that small levels of ambiguity do not significantly decrease the prediction accuracy. Finally, we showed that building the predictive model from ambiguous, rather than from statistically imputed high-resolution data, is a better approach to SNP-based prediction of HLA alleles. This is particularly true for non-European populations for which the HLA imputation error is larger than for a European population [36]. We showed that the loss of information due to the ambiguity in the training HLA data is smaller than the loss of information due to error incurred from HLA imputation.

As a future direction we intend to apply this approach on a larger sample and investigate whether the increased sample size can compensate for the loss of information due to the ambiguous HLA data. SNP-based prediction of HLA is highly dependent on the sample size because of the genetic diversity of HLA genes. Small samples do not have a good representation of the alleles in a population and tend to result in lower prediction accuracies. We believe that larger sample would compensate for the uncertainty in the data and result in a performance that would be comparable to the performance on the non-ambiguous data of the same size. Additionally, in order to avoid the errors within an allele family described in the previous section, we intend to implement a SNP selection step prior to building the model. This step would intelligently select the SNPs that are best able to distinguish alleles in the same allele family and, in general, as many alleles as possible in a validation data set.

Chapter 5

Conclusion and Discussion

This thesis addressed the challenges of prediction and analysis of HLA data. First, we defined a problem of prediction of HLA genes from SNP markers and described a two-step multi-label multi-class approach that utilizes strong correlations among labels to solve it [53, 54]. Second, we addressed the challenge of quantifying the uncertainty in HLA typing data, and demonstrated that the strong correlation structure is informative when it comes to resolving this uncertainty [52]. Third, we devised an approach to learning from uncertain HLA data in the SNP-based prediction of HLA genes that incorporates the uncertainty into the model [55]. In this chapter, we summarize the key results and provide an outlook of future work in this area.

5.1 Key Results

This section presents a summary of the major results that were produced as a part of this thesis.

5.1.1 Correlated multi-label multi-class prediction

We presented a novel approach for correlated multi-label multi-class prediction in the context of SNP-based prediction of HLA genes [53, 54]. Each HLA gene is a multi-class label. Labels are correlated because of the strong linkage disequilibrium in this region. We proposed a two step approach to predict HLA genes from SNPs that uses structural information that exists in the HLA region. In the first step we build label-independent

classifiers based on the Expectation-Maximization algorithm. In the second step we use HLA haplotype frequencies to enforce the correlation between labels and to make the final prediction. We demonstrated that adding haplotype frequency information increases the accuracy of prediction.

In addition, we evaluated the impact of HLA gene dependency on the prediction of HLA genes from SNP data [54]. We proposed integrating local and global dependencies among HLA genes into the prediction, and evaluated the impact of both approaches. Our results showed that the addition of global structure produces a more robust prediction with respect to the algorithm's parameter. However, given the small sample size of our data and a lack of population diversity, we expect that the local structure approach may have different impact in a larger data set or a data set of non-European origin. Large data sets are more likely to contain rare alleles, and strong local dependencies would be able to promote prediction ranking for these alleles. Additionally, haplotype patterns differ across populations. It has been observed that African haplotypes are shorter than non-African haplotypes, which implies that local structure should be more beneficial to prediction on data sets of African ancestry.

5.1.2 Quantifying uncertainty in HLA data

The high polymorphism in HLA presents a challenge when it comes to typing or sequencing HLA genes.The typing of HLA alleles is generally performed using DNA-based assays that are not always able to precisely identify alleles present. Exact allele-level HLA typing can be costly and laborious with the large and rapidly growing number of described HLA alleles. We described an information-theory based measure to quantify ambiguity content in an HLA typing [52]. We demonstrate that it can be objectively used to compare methods of HLA typing to each other in terms of the information they provide, in the context of each individual population. Our results show that intermediate-resolution single-pass sequence-based typing contains the least ambiguity and, therefore, the most certainty in allele prediction across all populations. In addition, we demonstrated the benefit of using haplotype frequencies in entropy calculations versus allele frequencies. When certain alleles occur together generally due to linkage disequilibrium between them, some ambiguity can be inferred away using this linkage information, which is inherently contained in haplotype frequencies. Our results show that using population

haplotype frequencies immensely reduces the ambiguity present in HLA typing.

5.1.3 Learning from uncertain data

We described a novel algorithm for prediction of uncertain data in the context of SNP-based prediction of HLA genes [55]. The existing uncertainty found in most HLA data sets today presents a challenge to current SNP-based methods for prediction of HLA alleles, since they require non-ambiguous data. While the typing methodology evolved over time and will evolve further, the majority of a stem-cell registry data still consists of ambiguous HLA assignments [20]. The Be The Match registry contains HLA typing of approximately 10 million donors as of this year, and the majority of these donors have ambiguous HLA data. In order to utilize the large amounts of historical HLA data in the Registry operations or research, we must develop algorithms that can successfully learn from uncertain data.

A way to obtain high-resolution HLA alleles from ambiguous data is to use one of the existing methods for resolving ambiguity from the ambiguous HLA data [36] and use the generated high-resolution HLA alleles in learning the SNP-based prediction method (Figure 4.1). However, depending on the levels of ambiguity in the HLA data, the errors generated by these methods can be prohibitively high. These errors are then further propagated when the incorrectly imputed HLA alleles are used to train the SNP-based prediction model. In this thesis we proposed a SNP-based approach for prediction of HLA alleles that utilizes ambiguous HLA genotypes directly as a learning source when building the predictive model. Additionally, we evaluated how different levels of ambiguity in HLA data impact the prediction performance and concluded that small levels of ambiguity do not significantly decrease the prediction accuracy. Finally, we showed that building the predictive model from ambiguous, rather than from statistically imputed high-resolution data, is generally a more accurate approach to SNP-based prediction of HLA alleles. This is particularly true for non-European populations for which the HLA imputation error is larger than for European population [36]. We showed that the loss of information due to the ambiguity in the training HLA data is smaller than the loss of information due to error incurred from HLA imputation.

5.2 Future Directions

We plan to continue our research in the following directions:

Investigation of computational approaches to correlated multi-label prediction: We intend to investigate possible improvements or alternative approaches to correlated multi-label multi-class prediction in the context of HLA genes. The algorithms described in this thesis include many parameters, and we intend to conduct a thorough experimental evaluation of how those parameters impact the overall performance of the method. Our current implementation of the base EM classifier is computationally expensive and cannot handle large numbers of SNPs. In the future, we plan to use a more scalable classifier that will handle larger numbers of SNPs and therefore make better initial predictions of HLA alleles. Another potential direction for improvement is the function to combine allele and haplotype scores into a final score. We currently use a weighted sum, however, we intend to investigate several other functions and their impact on prediction performance of the algorithm.

In this thesis we used haplotype frequencies published in [37] and generated from a relatively small sample of about $10K$ (compared to the current $\sim 10M$) individuals. In the future we intend to use more recently published haplotype frequencies generated from a much larger sample size ($\sim 6M$ individuals) and encompassing more populations (21 detailed populations) [20]. Using these frequencies may result in higher accuracies as they more adequately represent a population (e.g. using African Black population frequencies published in [20] may enhance predictions for YRI population used in this study because they match the YRI population more closely than African American frequencies published in [37]).

Finally, we intend to investigate different approaches to the implementation of gene-independent classifiers for the Step 1 of the algorithms described in Chapters 2 and 4. Recent methods for independent prediction of HLA genes from SNP data have demonstrated high accuracies [12, 85]. We plan to investigate how integrating some of these alternative methods into our framework impact the prediction performance.

Investigate the impact of the sample diversity: We plan to investigate the application of the proposed prediction framework on a data set of non-European origin. As demonstrated in the BC data set in Chapter 2, there is little room for improvement on large data sets of European origin, since high accuracies are obtained even without the addition of HLA haplotype information. However, data sets of non-European origin are less widely available, generally available at smaller sample size and have higher genetic diversity. As such, they present a bigger challenge when it comes to prediction of HLA genes and could greatly benefit from the gene dependency information. In addition, linkage disequilibrium patterns differ in different populations, and the addition of dependency information into the HLA prediction may be even more beneficial for a non-European population.

Additionally, we intend to investigate the impact of the sample size on the performance of our *Amb-EM* algorithm, in order to measure whether the increased sample size can compensate for the loss of information due to the ambiguous HLA data. SNP-based prediction of HLA is highly dependent on the sample size because of the genetic diversity of HLA genes. Small samples do not have a good representation of the alleles in a population and tend to result in lower prediction accuracies. We believe that a larger sample would compensate for the uncertainty in the data and result in a performance that would be comparable to the performance on the non-ambiguous data of the same size.

Prediction of haplotypes: An additional advantage of our approach for prediction of HLA alleles proposed in Chapter 2 is that it can be used for prediction of HLA haplotypes, rather than individual genes. Our algorithm can be extended to return a ranked list of HLA haplotypes that a new sample is likely to have. This is valuable in the context of stem cell transplant where donor search is performed based on several highest ranking haplotype pairs imputed from the ambiguous HLA data. It has also been shown that haplotype matching can inform on the graft-versus-host disease post transplant among HLA-identical transplant recipients [56]. We intend to extend our algorithm to predict phased HLA haplotypes for a new sample, and evaluate its performance on a sample with phased HLA data (such as the family data from the International Project HapMap [17]). Prediction of haplotypes rather than individual

HLA genes is more challenging due to the increased complexity of the problem.

Further investigation of data uncertainty: We plan to further investigate different approaches to learning from uncertain data. Our approach presented in Chapter 4 of this thesis incorporates all of the uncertainty present in the data into building the prediction model. We intend to investigate whether some of the uncertainty in the data can be resolved and removed before building the model without performance loss. The measure we introduced in Chapter 3 for quantifying the uncertainty in HLA data could be applied on uncertain data as the first step in attempt to remove the unlikely HLA uncertainties, and keep only the ones that will be informative for the prediction, rather than introduce noise.

Additionally, we aim to explore new ways for quantifying typing ambiguity that will be more intuitive and comparable across different system. We plan to differentiate between the content of ambiguity present in HLA genotypes and HLA un-phased genotypes, that has a more practical application in the context of donor and patient matching. Finally, we aim to analyze the HLA typing data in the Be The Match® registry using the improved ambiguity measure and demonstrate its value to other stem cell registries and HLA typing labs.

References

[1] Charu C Aggarwal, *On density based transforms for uncertain data mining*, Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on, IEEE, 2007, pp. 866–875.

[2] Charu C Aggarwal and Philip S Yu, *A survey of uncertain data algorithms and applications*, Knowledge and Data Engineering, IEEE Transactions on **21** (2009), no. 5, 609–623.

[3] E. Alvares Cherman, J. Metz, and M. Monard, *A simple approach to incorporate label dependency in multi-label classification*, Advances in Soft Computing (2010), 33–43.

[4] G Bentley, R Higuchi, B Hoglund, D Goodridge, D Sayer, EA Trachtenberg, and HA Erlich, *High-resolution, high-throughput hla genotyping by next-generation sequencing*, Tissue antigens **74** (2009), no. 5, 393–403.

[5] Werner Bochtler, Martin Maiers, Machteld Oudshoorn, Steven GE Marsh, Colette Raffoux, C Mueller, and Carolyn K Hurley, *World marrow donor association guidelines for use of hla nomenclature and its validation in the data exchange among hematopoietic stem cell donor registries and cord blood banks*, Bone marrow transplantation **39** (2007), no. 12, 737–741.

[6] Sharon R Browning and Brian L Browning, *Haplotype phasing: existing methods and new developments*, Nature Reviews Genetics **12** (2011), no. 10, 703–714.

[7] Paul R Burton, David G Clayton, Lon R Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P Kwiatkowski, Mark I McCarthy, Willem H Ouwehand, Nilesh J Samani, et al., *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls*, Nature **447** (2007), no. 7145, 661–678.

[8] Pedro Cano, *176-p measuring hla typing resolution*, Human Immunology **72** (2011), S129.

[9] Hurley CK, *DNA methods for HLA typing: A workbook for beginners*, ((1993, 1998, 2004, 2008)).

[10] Paul IW de Bakker and Soumya Raychaudhuri, *Interrogating the major histocompatibility complex with high-throughput genomics*, Human Molecular Genetics **21** (2012), no. R1, R29–R36.

[11] P.I.W. de Bakker, G. McVean, P.C. Sabeti, M.M. Miretti, T. Green, J. Marchini, X. Ke, A.J. Monsuur, P. Whittaker, M. Delgado, et al., *A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC*, Nature genetics **38** (2006), no. 10, 1166–1172.

[12] Alexander Dilthey, Stephen Leslie, Loukas Moutsianas, Judong Shen, Charles Cox, Matthew R Nelson, and Gil McVean, *Multi-population classical HLA type imputation*, PLOS Computational Biology **9** (2013), no. 2, e1002877.

[13] I. Evseeva, K.K. Nicodemus, C. Bonilla, S. Tonks, and W.F. Bodmer, *Linkage disequilibrium and age of HLA region SNPs in relation to classic HLA gene alleles within Europe*, European Journal of Human Genetics **18** (2010), no. 8, 924–932.

[14] M.M.A. Fernando, C.R. Stevens, E.C. Walsh, P.L. De Jager, P. Goyette, R.M. Plenge, T.J. Vyse, and J.D. Rioux, *Defining the role of the MHC in autoimmunity: a review and pooled analysis*, PLoS genetics **4** (2008), no. 4, e1000024.

[15] N. Ghamrawi and A. McCallum, *Collective multi-label classification*, Proceedings of the 14th ACM international conference on Information and knowledge management, ACM, 2005, pp. 195–200.

[16] Y. Ghodke, K. Joshi, A. Chopra, and B. Patwardhan, *HLA and disease*, European journal of epidemiology **20** (2005), no. 6, 475–488.

[17] Richard A Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, Fuli Yu, Huanming Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, et al., *The international hapmap project*, Nature **426** (2003), no. 6968, 789–796.

[18] S. Godbole and S. Sarawagi, *Discriminative methods for multi-labeled classification*, Advances in Knowledge Discovery and Data Mining (2004), 22–30.

[19] P.A. Gourraud, P. Lamiraux, N. El-Kadhi, C. Raffoux, and A. Cambon-Thomsen, *Inferred HLA haplotype information for donors from hematopoietic stem cells donor registries*, Human immunology **66** (2005), no. 5, 563–570.

[20] Loren Gragert, Abeer Madbouly, John Freeman, and Martin Maiers, *Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry*, Human immunology (2013).

[21] W Helmberg, J Hegland, CK Hurley, M Maiers, SGE Marsh, C Müller, and EH Rozemuller, *Going back to the roots: effective utilisation of HLA typing information for bone marrow registries requires full knowledge of the DNA sequences of the oligonucleotide reagents used in the testing*, Tissue Antigens **56** (2000), no. 1, 99–102.

[22] Wolfgang Helmberg, Douglas Hoffman, Michael Feolo, and Martin Maiers, *Population based characterisation of hla typing kits and ambiguous typing results: The frequency inferred typing (fit) index*, Human Immunology **66** (2005), no. 8, 36.

[23] Tomer Hertz and Chen Yanover, *Identifying hla supertypes by learning distance functions*, Bioinformatics **23** (2007), no. 2, e148–e155.

[24] R Holdsworth, CK Hurley, SGE Marsh, M Lau, HJ Noreen, JH Kempenich, M Setterholm, and M Maiers, *The hla dictionary 2008: a summary of hla-a,-b,-c,-drb1/3/4/5, and-dqb1 alleles and their association with serologically defined hla-a,-b,-c,-dr, and-dq antigens*, Tissue antigens **73** (2009), no. 2, 95–170.

[25] CK Hurley, M Maiers, J Ng, D Wagage, J Hegland, J Baisch, R Endres, M Fernandez-Vina, U Heine, S Hsu, et al., *Large-scale dna-based typing of hla-a and hla-b at low resolution is highly accurate specific and reliable*, Tissue Antigens **55** (2000), no. 4, 352–358.

[26] CK Hurley, M Setterholm, M Lau, MS Pollack, H Noreen, A Howard, M Fernandez-Vina, D Kukuruga, CR Müller, M Venance, et al., *Hematopoietic stem cell donor registry strategies for assigning search determinants and matching relationships*, Bone marrow transplantation **33** (2003), no. 4, 443–450.

[27] Xiaoming Jia, Buhm Han, Suna Onengut-Gumuscu, Wei-Min Chen, Patrick J Concannon, Stephen S Rich, Soumya Raychaudhuri, and Paul IW de Bakker, *Imputing amino acid polymorphisms in human leukocyte antigens*, PloS one **8** (2013), no. 6, e64683.

[28] J. Jiang, *Multi-label correlated semi-supervised learning for protein function prediction*, Bioinformatics Research and Applications (2011), 368–379.

[29] C. Kollman, M. Maiers, L. Gragert, C. Müller, M. Setterholm, M. Oudshoorn, and C.K. Hurley, *Estimation of HLA-A,-B,-DRB1 haplotype frequencies using mixed resolution data from a national registry with selective retyping of volunteers*, Human immunology **68** (2007), no. 12, 950–958.

[30] Charles J Krebs et al., *Ecological methodology*, vol. 620, Benjamin/Cummings Menlo Park, California, 1999.

[31] AM Lazaro, MA Fernandez-Viña, CJ Nulf, VM Fish, JE McGarry, CY Marcos, SN Miller, and P Stastny, *O215-optimization of high resolution dna typing of hla-a and b loci*, Human Immunology **47** (1996), no. 1, 43.

[32] S. Leslie, P. Donnelly, and G. McVean, *A statistical method for predicting classical HLA alleles from SNP data*, The American Journal of Human Genetics **82** (2008), no. 1, 48–56.

[33] Timothy R Lezon, Jayanth R Banavar, Marek Cieplak, Amos Maritan, and Nina V Fedoroff, *Using the principle of entropy maximization to infer genetic interaction*

networks from gene expression patterns, Proceedings of the National Academy of Sciences **103** (2006), no. 50, 19033–19038.

[34] S.S. Li, H. Wang, A. Smith, B. Zhang, X.C. Zhang, G. Schoch, D. Geraghty, J.A. Hansen, and L.P. Zhao, *Predicting multiallelic genes using unphased and flanking single nucleotide polymorphisms*, Genetic epidemiology **35** (2011), no. 2, 85–92.

[35] Jennifer Listgarten, Zabrina Brumme, Carl Kadie, Gao Xiaojiang, Bruce Walker, Mary Carrington, Philip Goulder, and David Heckerman, *Statistical resolution of ambiguous hla typing data*, PLoS computational biology **4** (2008), no. 2, e1000016.

[36] Abeer Madbouly, Loren Gragert, John Freeman, Nicole Leahy, Pierre-Antoine Gourraud, Jill Hollenbach, Malek Kamoun, Marcelo Fernandez-Vina, and Martin Maiers, *Validation of statistical imputation of allele-level multi-locus phased genotypes from ambiguous HLA assignments*, In print at Tissue antigens (2014).

[37] M. Maiers, L. Gragert, and W. Klitz, *High-resolution HLA alleles and haplotypes in the United States population*, Human immunology **68** (2007), no. 9, 779–788.

[38] M. Maiers, CK Hurley, L. Perlee, M. Fernandez-Vina, J. Baisch, D. Cook, P. Fraser, U. Heine, S. Hsu, MS Leffell, et al., *Maintaining updated DNA-based HLA assignments in the National Marrow Donor Program Bone Marrow Registry.*, Reviews in immunogenetics **2** (2000), no. 4, 449.

[39] M. Malkki, R. Single, M. Carrington, G. Thomson, and E. Petersdorf, *MHC microsatellite diversity and linkage disequilibrium among common HLA-A, HLA-B, DRB1 haplotypes: implications for unrelated donor hematopoietic transplantation and disease association studies*, Tissue Antigens **66** (2005), no. 2, 114–124.

[40] S.G.E. Marsh, E.D. Albert, W.F. Bodmer, R.E. Bontrop, B. Dupont, H.A. Erlich, D.E. Geraghty, J.A. Hansen, C.K. Hurley, B. Mach, et al., *Nomenclature for factors of the HLA system, 2004*, International journal of immunogenetics **32** (2005), no. 2, 107–159.

[41] Annalise M Martin, David Nolan, Silvana Gaudieri, Coral Ann Almeida, Richard Nolan, Ian James, Filipa Carvalho, Elizabeth Phillips, Frank T Christiansen, Anthony W Purcell, et al., *Predisposition to abacavir hypersensitivity conferred by*

HLA-B 5701 and a haplotypic Hsp70-Hom variant*, Proceedings of the National Academy of Sciences **101** (2004), no. 12, 4180–4185.

[42] Mark McCormack, Ana Alfirevic, Stephane Bourgeois, John J Farrell, Dalia Kasperavičiūtė, Mary Carrington, Graeme J Sills, Tony Marson, Xiaoming Jia, Paul IW de Bakker, et al., *HLA-A* 3101 and carbamazepine-induced hypersensitivity reactions in Europeans*, New England Journal of Medicine **364** (2011), no. 12, 1134–1143.

[43] X. Minzhu, L. Jing, and J. Tao, *Accurate HLA type inference using a weighted similarity graph*, BMC Bioinformatics **11**.

[44] Marcos M Miretti, Emily C Walsh, Xiayi Ke, Marcos Delgado, Mark Griffiths, Sarah Hunt, Jonathan Morrison, Pamela Whittaker, Eric S Lander, Lon R Cardon, et al., *A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms*, The American Journal of Human Genetics **76** (2005), no. 4, 634–646.

[45] Mehryar Mohri, *Learning from uncertain data*, Learning Theory and Kernel Machines, Springer, 2003, pp. 656–670.

[46] Yasuo Morishima, Takehiko Sasazuki, Hidetoshi Inoko, Takeo Juji, Tatsuya Akaza, Ken Yamamoto, Yoshihide Ishikawa, Shunichi Kato, Hiroshi Sao, Hisashi Sakamaki, et al., *The clinical significance of human leukocyte antigen (hla) allele compatibility in patients receiving a marrow transplant from serologically hla-a, hla-b, and hla-dr matched unrelated donors*, Blood **99** (2002), no. 11, 4200–4206.

[47] J Ng, CK Hurley, C Carter, LA Baxter-Lowe, D Bing, M Chopek, J Hegland, TD Lee, TC Li, S Hsu, et al., *Large-scale drb and dqb1 oligonucleotide typing for the nmdp registry: progress report from year 2*, Tissue Antigens **47** (1996), no. 1, 21–26.

[48] Jannifer Ng, Carolyn Katovich Nurlay, Lee Ann Baxter-Lowe, Michael Chepak, Patricia A Cappe, Janet Hagland, Debra KaKuraya, Dimitri Manes, Gayla Rosner, Barbara Schmeckpaper, et al., *Large-scale oligonucleotide typing for hla-drb1/3/4*

and hla-dqb1 is highly accurate, specific, and reliable, Tissue Antigens **42** (1993), no. 5, 473–479.

[49] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song, *Genotype and SNP calling from next-generation sequencing data*, Nature Reviews Genetics **12** (2011), no. 6, 443–451.

[50] G. Pandey, C. Myers, and V. Kumar, *Incorporating functional inter-relationships into protein function prediction algorithms*, BMC bioinformatics **10** (2009), no. 1, 142.

[51] Tan Pang-Ning, Michael Steinbach, Vipin Kumar, et al., *Introduction to data mining*, Library of Congress, 2006.

[52] Vanja Paunić, Loren Gragert, Abeer Madbouly, John Freeman, and Martin Maiers, *Measuring ambiguity in HLA typing methods*, PloS One **7** (2012), no. 8, e43585.

[53] Vanja Paunić, Michael Steinbach, Vipin Kumar, and Martin Maiers, *Prediction of HLA genes from SNP data and HLA haplotype frequencies*, Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on, IEEE, 2012, pp. 964–971.

[54] Vanja Paunić, Michael Steinbach, Abeer Madbouly, and Vipin Kumar, *Evaluation of label dependency for the prediction of HLA genes*, Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine, ACM, 2013.

[55] Vanja Paunic, Michael Steinbach, Abeer Madbouly, and Vipin Kumar, *Amb-EM: A SNP-based prediction of HLA alleles using ambiguous HLA data*, In preparation for ACM-BCB, 2014.

[56] Effie W Petersdorf, Mari Malkki, Ted A Gooley, Paul J Martin, and Zhen Guo, *Mhc haplotype matching for unrelated hematopoietic cell transplantation*, PLoS medicine **4** (2007), no. 1, e8.

[57] EW Petersdorf, Gary M Longton, C Anasetti, PJ Martin, EM Mickelson, AG Smith, and JA Hansen, *The significance of hla-drb1 matching on clinical outcome after hla-a, b, dr identical unrelated donor marrow transplantation*, Blood **86** (1995), no. 4, 1606–1613.

[58] Biao Qin, Yuni Xia, and Fang Li, *Dtu: a decision tree for uncertain data*, Advances in Knowledge Discovery and Data Mining, Springer, 2009, pp. 4–15.

[59] Biao Qin, Yuni Xia, Sunil Prabhakar, and Yicheng Tu, *A rule-based classification algorithm for uncertain data*, Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on, IEEE, 2009, pp. 1633–1640.

[60] J. Read, B. Pfahringer, G. Holmes, and E. Frank, *Classifier chains for multi-label classification*, Machine Learning and Knowledge Discovery in Databases (2009), 254–269.

[61] Jiangtao Ren, Sau Dan Lee, Xianlu Chen, Ben Kao, Reynold Cheng, and David Cheung, *Naive bayes classification of uncertain data*, Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on, IEEE, 2009, pp. 944–949.

[62] J. Robinson, A. Malik, P. Parham, JG Bodmer, and SGE Marsh, *IMGT/HLA database–a sequence database for the human major histocompatibility complex*, Tissue Antigens **55** (2000), no. 3, 280–287.

[63] J. Robinson, M.J. Waller, S.C. Fail, H. McWilliam, R. Lopez, P. Parham, and S.G.E. Marsh, *The IMGT/HLA database*, Nucleic acids research **37** (2009), no. suppl 1, D1013–D1017.

[64] James Robinson, Jason A Halliwell, Hamish McWilliam, Rodrigo Lopez, Peter Parham, and Steven GE Marsh, *The IMGT/HLA database*, Nucleic acids research **41** (2013), no. D1, D1222–D1227.

[65] Erik H Rozemuller and Marcel GJ Tilanus, *A computerized method to predict the discriminatory properties for class ii sequencing based typing*, Human immunology **46** (1996), no. 1, 27–34.

[66] Takehiko Sasazuki, Takeo Juji, Yasuo Morishima, Naoko Kinukawa, Hidehiko Kashiwabara, Hidetoshi Inoko, Takato Yoshida, Akinori Kimura, Tatsuya Akaza, Nobuhiro Kamikawaji, et al., *Effect of matching of class i hla alleles on clinical outcome after transplantation of hematopoietic stem cells from an unrelated donor*, New England Journal of Medicine **339** (1998), no. 17, 1177–1185.

[67] Thomas D Schneider, *A brief review of molecular information theory*, Nano communication networks **1** (2010), no. 3, 173–180.

[68] Stephan C Schuster, *Next-generation sequencing transforms todays biology*, Nature **200** (2007), no. 8.

[69] Debashis Sen and Sankar K Pal, *Generalized rough sets, entropy, and image ambiguity measures*, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on **39** (2009), no. 1, 117–128.

[70] M. Setty, A. Gusev, and I. Pe'er, *HLA type inference via haplotypes identical by descent*, Research in Computational Molecular Biology, Springer, 2010, pp. 491–505.

[71] Claude Elwood Shannon, *A mathematical theory of communication*, ACM SIGMOBILE Mobile Computing and Communications Review **5** (2001), no. 1, 3–55.

[72] T. Shiina, K. Hosomichi, H. Inoko, and J.K. Kulski, *The HLA genomic loci map: expression, interaction, diversity and disease*, Journal of human genetics **54** (2009), no. 1, 15–39.

[73] Christine F Skibola, Nicholas K Akers, Lucia Conde, Martha Ladner, Sharon K Hawbecker, Franziska Cohen, Fernanda Ribas, Henry A Erlich, Damian Goodridge, Elizabeth A Trachtenberg, et al., *Multi-locus hla class i and ii allele and haplotype associations with follicular lymphoma*, Tissue antigens **79** (2012), no. 4, 279–286.

[74] M.S. Sorower, *A literature survey on algorithms for multi-label learning.*

[75] Stephen R Spellman, Mary Eapen, Brent R Logan, Carlheinz Mueller, Pablo Rubinstein, Michelle I Setterholm, Ann E Woolfrey, Mary M Horowitz, Dennis L Confer, and Carolyn K Hurley, *A perspective on the selection of unrelated donors and cord blood units for transplantation*, Blood **120** (2012), no. 2, 259–265.

[76] Y. Tao, L. Sam, J. Li, C. Friedman, and Y.A. Lussier, *Information theory applied to the sparse gene ontology annotation network to predict novel gene function*, Bioinformatics **23** (2007), no. 13, i529–i538.

[77] G. Tsoumakas, I. Katakis, and I. Vlahavas, *Mining multi-label data*, Data Mining and Knowledge Discovery Handbook (2010), 667–685.

[78] G. Tsoumakas and I. Vlahavas, *Random k-labelsets: An ensemble method for multilabel classification*, Machine Learning: ECML 2007 (2007), 406–417.

[79] Jeffrey D Wall and Jonathan K Pritchard, *Haplotype blocks and linkage disequilibrium in the human genome*, Nature Reviews Genetics **4** (2003), no. 8, 587–597.

[80] Emily C Walsh, Kristie A Mather, Stephen F Schaffner, Lisa Farwell, Mark J Daly, Nick Patterson, Michael Cullen, Mary Carrington, Teodorica L Bugawan, Henry Erlich, et al., *An integrated haplotype map of the human major histocompatibility complex*, The American Journal of Human Genetics **73** (2003), no. 3, 580–590.

[81] Chunlin Wang, Sujatha Krishnakumar, Julie Wilhelmy, Farbod Babrzadeh, Lilit Stepanyan, Laura F Su, Douglas Levinson, Marcelo A Fernandez-Viña, Ronald W Davis, Mark M Davis, et al., *High-throughput, high-fidelity hla genotyping with deep sequencing*, Proceedings of the National Academy of Sciences **109** (2012), no. 22, 8676–8681.

[82] Jinbo Bi Tong Zhang, *Support vector classification with input data uncertainty*, Advances in neural information processing systems **17** (2005), 161–169.

[83] X.C. Zhang, B. Zhang, S.S. Li, X. Huang, J.A. Hansen, and L.P. Zhao, *Sequencing genes in silico using single nucleotide polymorphisms*, BMC genetics **13** (2012), no. 1, 6.

[84] Xinyi Zhang, Shuying Li, Hongwei Wang, John Hansen, and Lue Zhao, *Empirical evaluations of analytical issues arising from predicting HLA alleles using multiple SNPs*, BMC genetics **12** (2011), no. 1, 39.

[85] X Zheng, J Shen, C Cox, JC Wakefield, MG Ehm, MR Nelson, and BS Weir, *HIBAG - HLA genotype imputation with attribute bagging*, The Pharmacogenomics Journal (2013).